

Differential Equations

An Introduction for Engineers

by Matthew Charnley

August 6, 2021
(version Beta)

Typeset in L^AT_EX.

Copyright ©2021 Matthew Charnley



This work is dual licensed under the Creative Commons Attribution-Noncommercial-Share Alike 4.0 International License and the Creative Commons Attribution-Share Alike 4.0 International License. To view a copy of these licenses, visit <https://creativecommons.org/licenses/by-nc-sa/4.0/> or <https://creativecommons.org/licenses/by-sa/4.0/> or send a letter to Creative Commons PO Box 1866, Mountain View, CA 94042, USA.

You can use, print, duplicate, share this book as much as you want. You can base your own notes on it and reuse parts if you keep the license the same. You can assume the license is either the CC-BY-NC-SA or CC-BY-SA, whichever is compatible with what you wish to do, your derivative works must use at least one of the licenses. Derivative works must be prominently marked as such.

The date is the main identifier of version. The major version / edition number is raised only if there have been substantial changes.

See <https://sites.rutgers.edu/matthew.charnley> for more information (including contact information).

Contents

Introduction	11
0.1 Introduction to differential equations	11
0.2 Classification of differential equations	20
1 First Order Differential Equations	25
1.1 Integrals as solutions	25
1.2 Slope fields	30
1.3 Separable equations	34
1.4 Linear equations and the integrating factor	42
1.5 Existence and Uniqueness of Solutions	46
1.6 Numerical methods: Euler's method	53
1.7 Autonomous equations	60
1.8 Exact equations	72
1.9 Modeling with First Order Equations	82
1.10 Substitution	93
2 Higher order linear ODEs	99
2.1 Second order linear ODEs	99
2.2 Complex Roots and Euler's Formula	111
2.3 Repeated Roots and Reduction of Order	118
2.4 Mechanical vibrations	122
2.5 Nonhomogeneous equations	131
2.6 Forced oscillations and resonance	140
2.7 Higher order linear ODEs	149
3 Linear algebra	159
3.1 Vectors, mappings, and matrices	159
3.2 Matrix algebra	170
3.3 Elimination	182
3.4 Subspaces and dimension	191
3.5 Determinant	200
3.6 Eigenvalues and Eigenvectors	212
3.7 Kernel and Nullity	232

4 Systems of ODEs	237
4.1 Introduction to systems of ODEs	237
4.2 Matrices and linear systems	246
4.3 Linear systems of ODEs	257
4.4 Eigenvalue method	262
4.5 Two-dimensional systems and their vector fields	274
4.6 Nonhomogeneous systems	281
4.7 Second order systems and applications	291
4.8 Matrix exponentials	303
5 Nonlinear systems	313
5.1 Linearization, critical points, and equilibria	313
5.2 Stability and classification of isolated critical points	320
5.3 Applications of nonlinear systems	330
5.4 Limit cycles	343
5.5 Chaos	350
6 The Laplace transform	357
6.1 The Laplace transform	357
6.2 Transforms of derivatives and ODEs	364
6.3 Convolution	372
6.4 Dirac delta and impulse response	377
6.5 Solving PDEs with the Laplace transform	384
7 Power series methods	391
7.1 Power series	391
7.2 Series solutions of linear second order ODEs	400
7.3 Singular points and the method of Frobenius	407
8 Inner Products and Orthogonality	417
8.1 Inner product and projections	417
8.2 Special Matrices	425
8.3 Vector Spaces of Functions	440
9 Fourier series	447
9.1 Boundary value problems	447
9.2 The trigonometric series	456
9.3 More on the Fourier series	466
9.4 Sine and cosine series	476
9.5 Applications of Fourier series	484
10 Introduction to PDEs	491
10.1 First order linear PDE	491
10.2 Second order linear PDEs	498
10.3 The heat equation	507

CONTENTS	5
10.4 One-dimensional wave equation	520
10.5 D'Alembert solution of the wave equation	532
10.6 Steady state temperature and the Laplacian	538
10.7 Dirichlet problem in the circle and the Poisson kernel	544
11 Fourier transform	553
11.1 Fourier Integrals	553
11.2 Fourier Transform	562
A Introduction to MATLAB	569
A.1 The MATLAB Interface	569
A.2 Computation in MATLAB	571
A.3 Variables and Arrays	573
A.4 Functions and Anonymous Functions	575
A.5 Loops and Branching Statements	577
A.6 Plotting in MATLAB	578
A.7 Supplemental Code Files	580
B Prerequisite Material	591
B.1 Polynomials and Factoring	592
B.2 Complex Numbers	610
B.3 Differentiation and Integration Techniques	616
C Table of Laplace Transforms	627
Further Reading	629
Answers to Selected Exercises	631

Preface

Attributions

The main inspiration for this book, as well as the vast majority of the source material, is *Notes on Diffy Qs* by Jiří Lebl [JL]. The fact that the book is freely available and open-source provided the main motivation for creating this current text. It allowed this book to be put together in a timely manner to be useful. It significantly reduced the work needed to put together a free textbook that fit the course exactly.

Introduction to this Version

This text was originally designed for the Math 244 class at Rutgers University. This class is a first course in Differential Equations for Engineering majors. This class is taken immediately after Multivariable Calculus and does not assume any knowledge of linear algebra. Prior to the design of this book, the course used Boyce and DiPrima's *Elementary Differential Equations and Boundary Value Problems* [BD]. The course provided a very brief introduction to matrices in order to get to the information necessary to handle first order systems of differential equations. With the course being redesigned to include more linear algebra, I was pointed in the direction of Jiří Lebl's *Notes on Diffy Qs* [JL], which was meant to be a drop-in replacement for the Boyce and DiPrima text, and as of a more recent version of the text, contained an appendix on Linear Algebra.

In creating this book, I wanted to retain the style of *Notes on Diffy Qs* [JL] but shape the text into something that directly fit the course that we wanted to run. This included reorganizing some of the topics, extra contextualization of the concept of differential equations, sections devoted to modeling principles and how these equations can be derived, and guidance in using MATLAB to solve differential equations numerically. Specifically, the content added to this book is

- Appendix A that gives an introduction or review to coding in MATLAB, as well as references to sample MATLAB files that can be used to easily sketch slope fields and solution curves to differential equations.
- Section 1.9 on the accumulation equation, its use in mathematical models, and a discussion of parameter estimation, with inspiration taken from SIMIODE.
- Chapter 8 contains a discussion of orthogonality from vectors, to matrices, and then to function spaces, moving towards the idea of Fourier series.

- Appendix B on prerequisite material to be referred to when needed. Some of the material here was pulled from Stitz and Zeager's book *Precalculus* [SZ].
- Chapter 11 contains definitions and the basics of Fourier transforms in the context of solving partial differential equations, with some information adapted from [ZW].
- Exercises were added at the end of most sections of the text.

After designing this text for Math 244, only about half of the material in [JL] was used, and the rest of the material fit nicely with the next class in the sequence, Math 421. This class previously used *Advanced Engineering Mathematics* by Zill and Wright [ZW], and covered a smattering of topics throughout that book. The second half of this book was designed to put these topics in a single, freely-available text in the order that the course discussed them.

Acknowledgements

I would like to acknowledge David Molnar, who initially referred me to the *Notes on Diffy Qs* text [JL], as well as the *Precalculus* text [SZ], and provided inspiration and motivation to work on designing this text. For feedback during the development of the text, I want to acknowledge David Herrera, Yi-Zhi Huang, and many others who have helped over the development and refinement of this text. Finally, I want to acknowledge the Rutgers Open and Affordable Textbook Program for supporting the development and implementation of this text.

Introduction to *Notes on Diffy Qs*

This book [JL] originated from my class notes for Math 286 at the University of Illinois at Urbana-Champaign (UIUC) in Fall 2008 and Spring 2009. It is a first course on differential equations for engineers. Using this book, I also taught Math 285 at UIUC, Math 20D at University of California, San Diego (UCSD), and Math 4233 at Oklahoma State University (OSU). Normally these courses are taught with Edwards and Penney, *Differential Equations and Boundary Value Problems: Computing and Modeling* [EP], or Boyce and DiPrima's *Elementary Differential Equations and Boundary Value Problems* [BD], and this book aims to be more or less a drop-in replacement. Other books I used as sources of information and inspiration are E.L. Ince's classic (and inexpensive) *Ordinary Differential Equations* [I], Stanley Farlow's *Differential Equations and Their Applications* [F], now available from Dover, Berg and McGregor's *Elementary Partial Differential Equations* [BM], and William Trench's free book *Elementary Differential Equations with Boundary Value Problems* [T]. See the [Further Reading](#) chapter at the end of the book.

Computer resources

The book's website <https://www.jirka.org/diffyqs/> contains the following resources:

1. Interactive SAGE demos.
2. Online WeBWorK homeworks (using either your own WeBWorK installation or Edfinity) for most sections, customized for this book.
3. The PDFs of the figures used in this book.

I taught the UIUC courses using IODE (<https://faculty.math.illinois.edu/iode/>). IODE is a free software package that works with Matlab (proprietary) or Octave (free software). The graphs in the book were made with the Genius software (see <https://www.jirka.org/genius.html>). I use Genius in class to show these (and other) graphs.

Acknowledgments

Firstly, I would like to acknowledge Rick Laugesen. I used his handwritten class notes the first time I taught Math 286. My organization of this book through chapter 5, and the choice of material covered, is heavily influenced by his notes. Many examples and computations are taken from his notes. I am also heavily indebted to Rick for all the advice he has given me, not just on teaching Math 286. For spotting errors and other suggestions, I would also like to acknowledge (in no particular order): John P. D'Angelo, Sean Raleigh, Jessica Robinson, Michael Angelini, Leonardo Gomes, Jeff Winegar, Ian Simon, Thomas Wicklund, Eliot Brenner, Sean Robinson, Jannett Susberry, Dana Al-Quadi, Cesar Alvarez, Cem Bagdatlioglu, Nathan Wong, Alison Shive, Shawn White, Wing Yip Ho, Joanne Shin, Gladys Cruz, Jonathan Gomez, Janelle Louie, Navid Froutan, Grace Victorine, Paul Pearson, Jared Teague, Ziad Adwan, Martin Weilandt, Sönmez Şahutoğlu, Pete Peterson, Thomas Gresham, Prentiss Hyde, Jai Welch, Simon Tse, Andrew Browning, James Choi, Dusty

Grundmeier, John Marriott, Jim Kruidenier, Barry Conrad, Wesley Snider, Colton Koop, Sarah Morse, Erik Boczko, Asif Shakeel, Chris Peterson, Nicholas Hu, Paul Seeburger, Jonathan McCormick, David Leep, William Meisel, Shishir Agrawal, Tom Wan, Andres Valloud, and probably others I have forgotten. Finally, I would like to acknowledge NSF grants DMS-0900885 and DMS-1362337.

Introduction

0.1 Introduction to differential equations

Attribution: [JL], §0.2.

Learning Objectives

After this section, you will be able to:

- Identify a differential equation and determine the order of a differential equation,
- Verify that a function is a solution to a differential equation, and
- Solve some fundamental differential equations.

0.1.1 Differential equations

Consider the following situation:

An object falling through the air has its velocity affected by two factors: gravity and a drag force. The velocity downward is increased at a rate of $9.8m/s^2$ due to gravity, and it is decreased by a rate proportional to 0.3 times the current velocity of the object. If the ball is initially thrown downwards at a speed of $2m/s$, what will the velocity be 10 seconds later?

There might be enough information here to determine the velocity at any later point in time (it turns out, there is) but the information given isn't really about the velocity. Rather, information is given about the rate of change of the velocity. We know that the velocity will be increased at a rate of $9.8m/s^2$ due to gravity. How can this be interpreted? The rate of change has been discussed previously way back in Calculus 1; this is the derivative. Thus, the situation above gives information about the derivative of this unknown function for the velocity $v(t)$ of this object. Taking the two different factors into account, we can write an expression for this derivative, giving that

$$\frac{dv}{dt} = 9.8 - 0.3v.$$

Even though we don't know what $v(t)$ is, we know that it must affect the derivative of the velocity in this way, so we can write this equation. As it is an equation involving the

derivative of an unknown function $v(t)$, we call this a differential equation. Our goal here would be to use this information, plus the fact that the velocity at time zero is $v(0) = 2\text{m/s}$ to find the value of $v(10)$, or, more generally, the function $v(t)$ for any t .

The laws of physics, beyond just that of simple velocity, are generally written down as differential equations. Therefore, all of science and engineering use differential equations to some degree. Understanding differential equations is essential to understanding almost anything you will study in your science and engineering classes. You can think of mathematics as the language of science, and differential equations are one of the most important parts of this language as far as science and engineering are concerned. As an analogy, suppose all your classes from now on were given half in Swahili and half in English. It would be important to first learn Swahili, or you would have a very tough time getting a good grade in your classes. Without it, you might be able to make sense of some of what is going on, but would definitely be missing an important part of the picture.

Definition 0.1.1

A *differential equation* is an equation that involves one or more derivatives of an unknown function. For a differential equation, the *order* of the differential equation is the highest order derivative that appears in the equation.

One example of a first order differential equation is

$$\frac{dx}{dt} + x = 2 \cos t. \quad (1)$$

Here x is the *dependent variable* and t is the *independent variable*. Note that we can use any letter we want for the dependent and independent variables. This equation arises from Newton's law of cooling where the ambient temperature oscillates with time.

To make sure that everything is well-defined, we will assume that we can always write our differential equation with the highest order derivative written as a function of all lower derivatives and the independent variable. For the previous example, since we can write (1) as

$$\frac{dx}{dt} = 2 \cos t - x$$

where the highest derivative x' is written as a function of t and x , we have a proper differential equation. On the other hand, something like

$$\left(\frac{dy}{dt}\right)^2 + y^2 = 1 \quad (2)$$

is not a proper differential equation because we can't solve for $\frac{dy}{dt}$. This expression could either be written as

$$\frac{dy}{dt} = \sqrt{1 - y^2} \quad \text{or} \quad \frac{dy}{dt} = -\sqrt{1 - y^2},$$

and while both of these are proper differential equations, the version in (2) is not.

For some equations, like $y' = y^2$, the independent variable is not explicitly stated. We could be looking for a function $y(t)$ or a function $y(x)$ (or y of any other variable) and without

any other information, any of these is correct. Usually, there will be information in the problem statement to indicate that the independent variable is something like time, in which case everything should be written in terms of t . It is for this reason that Leibniz notation is preferred for derivatives; an equation like

$$\frac{dy}{dt} = y^2$$

is unambiguously looking for any answer $y(t)$.

Example 0.1.1: All of the below are differential equations

$$\frac{dy}{dt} = e^t y \quad y'' + y^2 = t \sin y$$

$$\frac{d^4y}{dx^4} - 3x \frac{d^2y}{dx^2} = x \quad y''' + (y'')^2 - 3y = t^4.$$

Note that any letter can be used for the unknown function and its dependent variable. The order of these equations are 1, 2, 4, and 3 respectively.

0.1.2 Solutions of differential equations

Solving the differential equation means finding the function that, when we plug it into the differential equation, gives a true statement. For example, take (1) from the previous section. In this case, this means that we want to find a function of t , which we call x , such that when we plug x , t , and $\frac{dx}{dt}$ into (1), the equation holds; that is, the left hand side equals the right hand side. It is the same idea as it would be for a normal (algebraic) equation of just x and t . We claim that

$$x = x(t) = \cos t + \sin t$$

is a *solution*. How do we check? We simply plug x into equation (1)! First we need to compute $\frac{dx}{dt}$. We find that $\frac{dx}{dt} = -\sin t + \cos t$. Now let us compute the left-hand side of (1).

$$\frac{dx}{dt} + x = \underbrace{(-\sin t + \cos t)}_{\frac{dx}{dt}} + \underbrace{(\cos t + \sin t)}_x = 2 \cos t.$$

Yay! We got precisely the right-hand side. But there is more! We claim $x = \cos t + \sin t + e^{-t}$ is also a solution. Let us try,

$$\frac{dx}{dt} = -\sin t + \cos t - e^{-t}.$$

We plug into the left-hand side of (1)

$$\frac{dx}{dt} + x = \underbrace{(-\sin t + \cos t - e^{-t})}_{\frac{dx}{dt}} + \underbrace{(\cos t + \sin t + e^{-t})}_x = 2 \cos t.$$

And it works yet again!

So there can be many different solutions. For this equation all solutions can be written in the form

$$x = \cos t + \sin t + Ce^{-t},$$

for some constant C . Different constants C will give different solutions, so there are really infinitely many possible solutions. See [Figure 1](#) for the graph of a few of these solutions. We will see how we find these solutions a few lectures from now.

Solving differential equations can be quite hard. There is no general method that solves every differential equation. We will generally focus on how to get exact formulas for solutions of certain differential equations, but we will also spend a little bit of time on getting approximate solutions. And we will spend some time on understanding the equations without solving them.

Most of this book is dedicated to *ordinary differential equations* or ODEs, that is, equations with only one independent variable, where derivatives are only with respect to this one variable. If there are several independent variables, we get *partial differential equations* or PDEs.

Even for ODEs, which are very well understood, it is not a simple question of turning a crank to get answers. When you can find exact solutions, they are usually preferable to approximate solutions. It is important to understand how such solutions are found. Although in real applications you will leave much of the actual calculations to computers, you need to understand what they are doing. It is often necessary to simplify or transform your equations into something that a computer can understand and solve. You may even need to make certain assumptions and changes in your model to achieve this.

To be a successful engineer or scientist, you will be required to solve problems in your job that you have never seen before. It is important to learn problem solving techniques, so that you may apply those techniques to new problems. A common mistake is to expect to learn some prescription for solving all the problems you will encounter in your later career. This course is no exception.

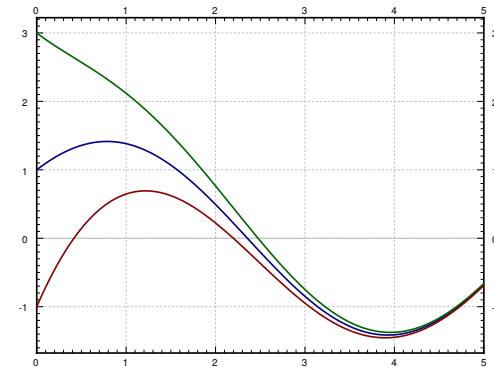


Figure 1: Few solutions of $\frac{dx}{dt} + x = 2 \cos t$.

0.1.3 Differential equations in practice

So how do we use differential equations in science and engineering? The main way this takes place is through the process of mathematical modeling. First, we have some *real-world problem* we wish to understand. We make some simplifying assumptions and create a *mathematical model*, which is a translation of this real-world problem into a set of differential equations. Think back to the example at the beginning of this section. We took a physical situation (a falling object) with some knowledge about how it behaves and turned that into a differential equation that describes the velocity over time. Then we apply mathematics to get some sort of a *mathematical solution*. Finally, we need to interpret our results, determining what this mathematical solution says about the real-world problem we started with. For instance, in the example at the start of the section, we could find the function $v(t)$, but then need to interpret that if we were to plug 10 into this function, we will get the velocity 10 seconds later.

Learning how to formulate the mathematical model and how to interpret the results is what your physics and engineering classes do. In this course, we will focus mostly on the mathematical analysis. This will be interspersed with discussions of this modeling process to give some context to what we are doing, and give practice for what will be seen in future physics and engineering classes.

Let us look at an example of this process. One of the most basic differential equations is the standard *exponential growth model*. Let P denote the population of some bacteria on a Petri dish. We assume that there is enough food and enough space. Then the rate of growth of bacteria is proportional to the population—a large population grows quicker. Let t denote time (say in seconds) and P the population. Our model is

$$\frac{dP}{dt} = kP,$$

for some positive constant $k > 0$.

Example 0.1.2: Suppose there are 100 bacteria at time 0 and 200 bacteria 10 seconds later. How many bacteria will there be 1 minute from time 0 (in 60 seconds)?

Solution: First we need to solve the equation. We claim that a solution is given by

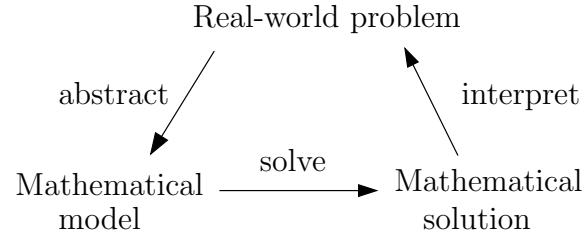
$$P(t) = Ce^{kt},$$

where C is a constant. Let us try:

$$\frac{dP}{dt} = Cke^{kt} = kP.$$

And it really is a solution.

OK, now what? We do not know C , and we do not know k . But we know something. We know $P(0) = 100$, and we know



$P(10) = 200$. Let us plug these conditions in and see what happens.

$$\begin{aligned} 100 &= P(0) = Ce^{k0} = C, \\ 200 &= P(10) = 100e^{k10}. \end{aligned}$$

Therefore, $2 = e^{10k}$ or $\frac{\ln 2}{10} = k \approx 0.069$. So

$$P(t) = 100e^{(\ln 2)t/10} \approx 100e^{0.069t}.$$

At one minute, $t = 60$, the population is $P(60) = 6400$. See [Figure 2](#).

Let us talk about the interpretation of the results. Does our solution mean that there must be exactly 6400 bacteria on the plate at 60s? No! We made assumptions that might not be true exactly, just approximately. If our assumptions are reasonable, then there will be approximately 6400 bacteria. Also, in real life P is a discrete quantity, not a real number. However, our model has no problem saying that for example at 61 seconds, $P(61) \approx 6859.35$.

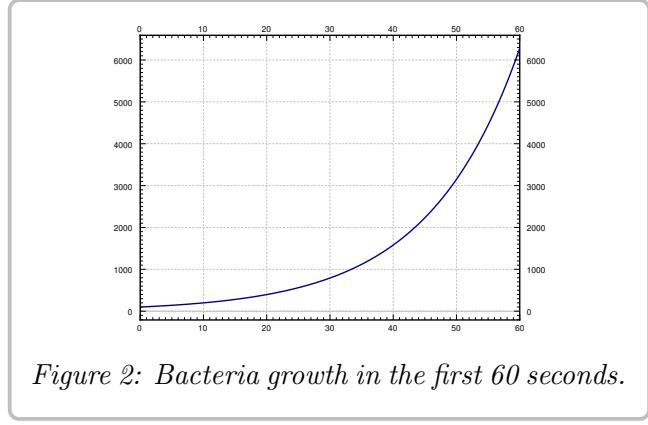


Figure 2: Bacteria growth in the first 60 seconds.

Normally, the k in $P' = kP$ is known, and we want to solve the equation for different *initial conditions*. What does that mean? Take $k = 1$ for simplicity. Suppose we want to solve the equation $\frac{dP}{dt} = P$ subject to $P(0) = 1000$ (the initial condition). Then the solution turns out to be (exercise)

$$P(t) = 1000e^t.$$

We call $P(t) = Ce^t$ the *general solution*, as every solution of the equation can be written in this form for some constant C . We need an initial condition to find out what C is, in order to find the *particular solution* we are looking for. Generally, when we say “particular solution,” we just mean some solution.

0.1.4 Four fundamental equations

A few equations appear often and it is useful to just memorize what their solutions are. Let us call them the four fundamental equations. Their solutions are reasonably easy to guess by recalling properties of exponentials, sines, and cosines. They are also simple to check, which is something that you should always do. No need to wonder if you remembered the solution correctly.

First such equation is

$$\frac{dy}{dx} = ky,$$

for some constant $k > 0$. Here y is the dependent and x the independent variable. The general solution for this equation is

$$y(x) = Ce^{kx}.$$

We saw above that this function is a solution, although we used different variable names.

Next,

$$\frac{dy}{dx} = -ky,$$

for some constant $k > 0$. The general solution for this equation is

$$y(x) = Ce^{-kx}.$$

Exercise 0.1.1: Check that the y given is really a solution to the equation.

Next, take the *second order differential equation*

$$\frac{d^2y}{dx^2} = -k^2y,$$

for some constant $k > 0$. The general solution for this equation is

$$y(x) = C_1 \cos(kx) + C_2 \sin(kx).$$

Since the equation is a second order differential equation, we have two constants in our general solution.

Exercise 0.1.2: Check that the y given is really a solution to the equation.

Finally, consider the second order differential equation

$$\frac{d^2y}{dx^2} = k^2y,$$

for some constant $k > 0$. The general solution for this equation is

$$y(x) = C_1 e^{kx} + C_2 e^{-kx},$$

or

$$y(x) = D_1 \cosh(kx) + D_2 \sinh(kx).$$

For those that do not know, \cosh and \sinh are defined by

$$\cosh x = \frac{e^x + e^{-x}}{2}, \quad \sinh x = \frac{e^x - e^{-x}}{2}.$$

They are called the *hyperbolic cosine* and *hyperbolic sine*. These functions are sometimes easier to work with than exponentials. They have some nice familiar properties such as $\cosh 0 = 1$, $\sinh 0 = 0$, and $\frac{d}{dx} \cosh x = \sinh x$ (no that is not a typo) and $\frac{d}{dx} \sinh x = \cosh x$.

Exercise 0.1.3: Check that both forms of the y given are really solutions to the equation.

Example 0.1.3: In equations of higher order, you get more constants you must solve for to get a particular solution. The equation $\frac{d^2y}{dx^2} = 0$ has the general solution $y = C_1x + C_2$; simply integrate twice and don't forget about the constant of integration. Consider the initial conditions $y(0) = 2$ and $y'(0) = 3$. We plug in our general solution and solve for the constants:

$$2 = y(0) = C_1 \cdot 0 + C_2 = C_2, \quad 3 = y'(0) = C_1.$$

In other words, $y = 3x + 2$ is the particular solution we seek.

0.1.5 Exercises

*Note: Exercises marked with * have answers in the back of the book.*

Exercise 0.1.4: Show that $x = e^{4t}$ is a solution to $x''' - 12x'' + 48x' - 64x = 0$.

Exercise 0.1.5:* Show that $x = e^{-2t}$ is a solution to $x'' + 4x' + 4x = 0$.

Exercise 0.1.6: Show that $x = e^t$ is not a solution to $x''' - 12x'' + 48x' - 64x = 0$.

Exercise 0.1.7: Is $y = \sin t$ a solution to $(\frac{dy}{dt})^2 = 1 - y^2$? Justify.

Exercise 0.1.8:* Is $y = x^2$ a solution to $x^2y'' - 2y = 0$? Justify.

Exercise 0.1.9: Let $y'' + 2y' - 8y = 0$. Now try a solution of the form $y = e^{rx}$ for some (unknown) constant r . Is this a solution for some r ? If so, find all such r .

Exercise 0.1.10:* Let $xy'' - y' = 0$. Try a solution of the form $y = x^r$. Is this a solution for some r ? If so, find all such r .

Exercise 0.1.11: Verify that $x = Ce^{-2t}$ is a solution to $x' = -2x$. Find C to solve for the initial condition $x(0) = 100$.

Exercise 0.1.12: Verify that $x = C_1e^{-t} + C_2e^{2t}$ is a solution to $x'' - x' - 2x = 0$. Find C_1 and C_2 to solve for the initial conditions $x(0) = 10$ and $x'(0) = 0$.

Exercise 0.1.13:* Verify that $x = C_1e^t + C_2$ is a solution to $x'' - x' = 0$. Find C_1 and C_2 so that x satisfies $x(0) = 10$ and $x'(0) = 100$.

Exercise 0.1.14: Find a solution to $(x')^2 + x^2 = 4$ using your knowledge of derivatives of functions that you know from basic calculus.

Exercise 0.1.15:* Solve $\frac{d\varphi}{ds} = 8\varphi$ and $\varphi(0) = -9$.

Exercise 0.1.16: Solve:

a) $\frac{dA}{dt} = -10A, \quad A(0) = 5$

b) $\frac{dH}{dx} = 3H, \quad H(0) = 1$

c) $\frac{d^2y}{dx^2} = 4y, \quad y(0) = 0, \quad y'(0) = 1$

d) $\frac{d^2x}{dy^2} = -9x, \quad x(0) = 1, \quad x'(0) = 0$

Exercise 0.1.17:* Solve:

a) $\frac{dx}{dt} = -4x, \quad x(0) = 9$

b) $\frac{d^2x}{dt^2} = -4x, \quad x(0) = 1, \quad x'(0) = 2$

c) $\frac{dp}{dq} = 3p, \quad p(0) = 4$

d) $\frac{d^2T}{dx^2} = 4T, \quad T(0) = 0, \quad T'(0) = 6$

Exercise 0.1.18: Is there a solution to $y' = y$, such that $y(0) = y(1)$?

Exercise 0.1.19: The population of city X was 100 thousand 20 years ago, and the population of city X was 120 thousand 10 years ago. Assuming constant growth, you can use the exponential population model (like for the bacteria). What do you estimate the population is now?

Exercise 0.1.20: Suppose that a football coach gets a salary of one million dollars now, and a raise of 10% every year (so exponential model, like population of bacteria). Let s be the salary in millions of dollars, and t is time in years.

a) What is $s(0)$ and $s(1)$.

b) Approximately how many years will it take for the salary to be 10 million.

c) Approximately how many years will it take for the salary to be 20 million.

d) Approximately how many years will it take for the salary to be 30 million.

0.2 Classification of differential equations

Attribution: [JL], §0.3.

Learning Objectives

After this section, you will be able to:

- Classify equation as ordinary or partial differential equations,
- Identify whether an equation is linear or non-linear, and
- Classify linear equations as homogenous, non-homogenous, or constant coefficient, as appropriate.

There are many types of differential equations, and we classify them into different categories based on their properties. Let us quickly go over the most basic classification. We already saw the distinction between ordinary and partial differential equations:

Definition 0.2.1

- *Ordinary differential equations* or (ODE) are equations where the derivatives are taken with respect to only one variable. That is, there is only one independent variable.
- *Partial differential equations* or (PDE) are equations that depend on partial derivatives of several variables. That is, there are several independent variables.

Let us see some examples of ordinary differential equations:

$$\frac{dy}{dt} = ky, \quad (\text{Exponential growth})$$

$$\frac{dy}{dt} = k(A - y), \quad (\text{Newton's law of cooling})$$

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = f(t). \quad (\text{Mechanical vibrations})$$

And of partial differential equations:

$$\frac{\partial y}{\partial t} + c\frac{\partial y}{\partial x} = 0, \quad (\text{Transport equation})$$

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \quad (\text{Heat equation})$$

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}. \quad (\text{Wave equation in 2 dimensions})$$

If there are several equations working together, we have a so-called *system of differential equations*. For example,

$$y' = x, \quad x' = y$$

is a simple system of ordinary differential equations. Maxwell's equations for electromagnetics,

$$\begin{aligned}\nabla \cdot \vec{D} &= \rho, & \nabla \cdot \vec{B} &= 0, \\ \nabla \times \vec{E} &= -\frac{\partial \vec{B}}{\partial t}, & \nabla \times \vec{H} &= \vec{J} + \frac{\partial \vec{D}}{\partial t},\end{aligned}$$

are a system of partial differential equations. The divergence operator $\nabla \cdot$ and the curl operator $\nabla \times$ can be written out in partial derivatives of the functions involved in the x , y , and z variables.

In the first chapter, we will start attacking first order ordinary differential equations, that is, equations of the form $\frac{dy}{dx} = f(x, y)$. In general, lower order equations are easier to work with and have simpler behavior, which is why we start with them.

We also distinguish how the dependent variables appear in the equation (or system). In particular, we say an equation is *linear* if the dependent variable (or variables) and their derivatives appear linearly, that is only as first powers, they are not multiplied together, and no other functions of the dependent variables appear. In other words, the equation is a sum of terms, where each term is some function of the independent variables or some function of the independent variables multiplied by a dependent variable or its derivative. Otherwise, the equation is called *nonlinear*. For example, an ordinary differential equation is linear if it can be put into the form

$$a_n(x) \frac{d^n y}{dx^n} + a_{n-1}(x) \frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_1(x) \frac{dy}{dx} + a_0(x)y = b(x). \quad (3)$$

The functions a_0, a_1, \dots, a_n are called the *coefficients*. The equation is allowed to depend arbitrarily on the independent variable. So

$$e^x \frac{d^2 y}{dx^2} + \sin(x) \frac{dy}{dx} + x^2 y = \frac{1}{x} \quad (4)$$

is still a linear equation as y and its derivatives only appear linearly. The equation

$$\cos(x) \frac{d^2 y}{dx^2} - xy + \frac{e^x}{x} = 0$$

is also linear, even though it is not initially in the correct form. From this equation, we can move the last term over to the right-hand side as a $-\frac{e^x}{x}$, and then it is in the correct form, with the $\frac{dy}{dt}$ term missing (or has coefficient zero).

All the equations and systems above as examples are linear. It may not be immediately obvious for Maxwell's equations unless you write out the divergence and curl in terms of partial derivatives. Let us see some nonlinear equations. For example Burger's equation,

$$\frac{\partial y}{\partial t} + y \frac{\partial y}{\partial x} = \nu \frac{\partial^2 y}{\partial x^2},$$

is a nonlinear second order partial differential equation. It is nonlinear because y and $\frac{\partial y}{\partial x}$ are multiplied together. The equation

$$\frac{dx}{dt} = x^2 \quad (5)$$

is a nonlinear first order differential equation as there is a second power of the dependent variable x .

Definition 0.2.2

A linear equation may further be called *homogeneous* if all terms depend on the dependent variable. That is, if no term is a function of the independent variables alone. Otherwise, the equation is called *nonhomogeneous* or *inhomogeneous*.

For example, the exponential growth equation, the wave equation, or the transport equation above are homogeneous. The mechanical vibrations equation above is nonhomogeneous as long as $f(t)$ is not the zero function. Similarly, if the ambient temperature A is nonzero, Newton's law of cooling is nonhomogeneous. A homogeneous linear ODE can be put into the form

$$a_n(x) \frac{d^n y}{dx^n} + a_{n-1}(x) \frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_1(x) \frac{dy}{dx} + a_0(x)y = 0.$$

Compare to (3) and notice there is no function $b(x)$.

If the coefficients of a linear equation are actually constant functions, then the equation is said to have *constant coefficients*. The coefficients are the functions multiplying the dependent variable(s) or one of its derivatives, not the function $b(x)$ standing alone. A constant coefficient nonhomogeneous ODE is an equation of the form

$$a_n \frac{d^n y}{dx^n} + a_{n-1} \frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_1 \frac{dy}{dx} + a_0 y = b(x),$$

where a_0, a_1, \dots, a_n are all constants, but b may depend on the independent variable x . The mechanical vibrations equation above is a constant coefficient nonhomogeneous second order ODE. The same nomenclature applies to PDEs, so the transport equation, heat equation and wave equation are all examples of constant coefficient linear PDEs.

Finally, an equation (or system) is called *autonomous* if the equation does not explicitly depend on the independent variable. For autonomous ordinary differential equations, the independent variable is then thought of as time. Autonomous equation means an equation that does not change with time. For example, Newton's law of cooling is autonomous, so is equation (5). On the other hand, mechanical vibrations or (4) are not autonomous.

0.2.1 Exercises

Exercise 0.2.1: Classify the following equations. Are they ODE or PDE? Is it an equation or a system? What is the order? Is it linear or nonlinear, and if it is linear, is it homogeneous, constant coefficient? If it is an ODE, is it autonomous?

- | | |
|--|--|
| a) $\sin(t) \frac{d^2 x}{dt^2} + \cos(t)x = t^2$ | b) $\frac{\partial u}{\partial x} + 3 \frac{\partial u}{\partial y} = xy$ |
| c) $y'' + 3y + 5x = 0, \quad x'' + x - y = 0$ | d) $\frac{\partial^2 u}{\partial t^2} + u \frac{\partial^2 u}{\partial s^2} = 0$ |
| e) $x'' + tx^2 = t$ | f) $\frac{d^4 x}{dt^4} = 0$ |

Exercise 0.2.2:* Classify the following equations. Are they ODE or PDE? Is it an equation or a system? What is the order? Is it linear or nonlinear, and if it is linear, is it homogeneous, constant coefficient? If it is an ODE, is it autonomous?

a) $\frac{\partial^2 v}{\partial x^2} + 3\frac{\partial^2 v}{\partial y^2} = \sin(x)$

b) $\frac{dx}{dt} + \cos(t)x = t^2 + t + 1$

c) $\frac{d^7 F}{dx^7} = 3F(x)$

d) $y'' + 8y' = 1$

e) $x'' + tyx' = 0, \quad y'' + txy = 0$

f) $\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial s^2} + u^2$

Exercise 0.2.3: If $\vec{u} = (u_1, u_2, u_3)$ is a vector, we have the divergence $\nabla \cdot \vec{u} = \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} + \frac{\partial u_3}{\partial z}$ and curl $\nabla \times \vec{u} = \left(\frac{\partial u_3}{\partial y} - \frac{\partial u_2}{\partial z}, \frac{\partial u_1}{\partial z} - \frac{\partial u_3}{\partial x}, \frac{\partial u_2}{\partial x} - \frac{\partial u_1}{\partial y} \right)$. Notice that curl of a vector is still a vector. Write out Maxwell's equations in terms of partial derivatives and classify the system.

Exercise 0.2.4: Suppose F is a linear function, that is, $F(x, y) = ax + by$ for constants a and b . What is the classification of equations of the form $F(y', y) = 0$.

Exercise 0.2.5: Write down an explicit example of a third order, linear, nonconstant coefficient, nonautonomous, nonhomogeneous system of two ODE such that every derivative that could appear, does appear.

Exercise 0.2.6:* Write down the general zeroth order linear ordinary differential equation. Write down the general solution.

Exercise 0.2.7:* For which k is $\frac{dx}{dt} + x^k = t^{k+2}$ linear. Hint: there are two answers.

Exercise 0.2.8: Write out an explicit example of a non-homogeneous fourth order, linear, constant coefficient differential equation. where all possible derivatives of the unknown function y appear.

Chapter 1

First Order Differential Equations

In this chapter, we begin by discussing first order differential equations. As they have the lowest possible order, only containing one derivative of the unknown function, they tend to be the simplest equations to try to analyze and solve. This doesn't mean that we'll be able to solve all of them, but we can make a decent effort at a fair number of them. These equations are also very common in modeling problems, as most principles from science and engineering give us a way to express the rate of change of a given quantity. This setup gives rise to a first order differential equation involving that quantity, which, if we can solve it, will tell us how the quantity evolves over time. Even if we can't solve the equation analytically, a numerical solution may allow us to predict the behavior of a system over time and design it to best fit our needs.

1.1 Integrals as solutions

Attribution: [JL], §1.1.

Learning Objectives

After this section, you will be able to:

- Solve a first order differential equation by direct integration and
- Understand the difference between a general solution and the solution to an initial value problem.

A first order ODE is an equation of the form

$$\frac{dy}{dx} = f(x, y),$$

or just

$$y' = f(x, y).$$

Some examples that fit this form are

$$y' = x^2y - e^x \sin x$$

and

$$y' = e^y(x^2 + 1) - \cos(y).$$

Looking back at the last section, the first of these is linear and the second is not. In general, there is no simple formula or procedure one can follow to find solutions. In the next few sections we will look at special cases where solutions are not difficult to obtain. In this section, let us assume that f is a function of x alone, that is, the equation is

$$y' = f(x). \quad (1.1)$$

We could just integrate (antidifferentiate) both sides with respect to x .

$$\int y'(x) dx = \int f(x) dx + C,$$

that is

$$y(x) = \int f(x) dx + C.$$

This $y(x)$ is actually the general solution. So to solve (1.1), we find some antiderivative of $f(x)$ and then we add an arbitrary constant to get the general solution.

Now is a good time to discuss a point about calculus notation and terminology. Calculus textbooks muddy the waters by talking about the integral as primarily the so-called indefinite integral. The indefinite integral is really the *antiderivative* (in fact the whole one-parameter family of antiderivatives). There really exists only one integral and that is the definite integral. The only reason for the indefinite integral notation is that we can always write an antiderivative as a (definite) integral. That is, by the fundamental theorem of calculus we can always write $\int f(x) dx + C$ as

$$\int_{x_0}^x f(t) dt + C.$$

Hence the terminology *to integrate* when we may really mean *to antidifferentiate*. Integration is just one way to compute the antiderivative (and it is a way that always works, see the following examples). Integration is defined as the area under the graph and it also happens to also compute antiderivatives. For sake of consistency, we will keep using the indefinite integral notation when we want an antiderivative, and you should *always* think of the definite integral as a way to write it.

Example 1.1.1: Find the general solution of $y' = 3x^2$.

Solution: Elementary calculus tells us that the general solution must be $y = x^3 + C$. Let us check by differentiating: $y' = 3x^2$. We got *precisely* our equation back. □

Normally, we will also have an initial condition such as $y(x_0) = y_0$ for some two numbers x_0 and y_0 (x_0 is often 0, but not always). If we do, the combination of a differential equation and an initial condition is called an initial value problem. We can then write the solution as a definite integral in a nice way. Suppose our problem is $y' = f(x)$, $y(x_0) = y_0$. Then the solution is

$$y(x) = \int_{x_0}^x f(s) ds + y_0. \quad (1.2)$$

Let us check! We compute $y' = f(x)$, via the fundamental theorem of calculus, and by Jupiter, y is a solution. Is it the one satisfying the initial condition? Well, $y(x_0) = \int_{x_0}^{x_0} f(x) dx + y_0 = y_0$. It is!

Do note that the definite integral and the indefinite integral (antidifferentiation) are completely different beasts. The definite integral always evaluates to a number. Therefore, (1.2) is a formula we can plug into the calculator or a computer, and it will be happy to calculate specific values for us. We will easily be able to plot the solution and work with it just like with any other function. It is not so crucial to always find a closed form for the antiderivative.

Example 1.1.2: Solve

$$y' = e^{-x^2}, \quad y(0) = 1.$$

Solution: By the preceding discussion, the solution must be

$$y(x) = \int_0^x e^{-s^2} ds + 1.$$

Here is a good way to make fun of your friends taking second semester calculus. Tell them to find the closed form solution. Ha ha ha (bad math joke). It is not possible (in closed form). There is absolutely nothing wrong with writing the solution as a definite integral. This particular integral is in fact very important in statistics. \square

While there is nothing wrong with writing solutions as a definite integral, they should be simplified and evaluated if possible. Given the differential equation

$$y' = 3x^2, \quad y(2) = 6,$$

the solution can be written as

$$y(x) = \int_2^x 3s^2 ds + 6.$$

However, it is much more convenient, both for human reasoning and computers, to write this solution as

$$y(x) = x^3 - 2.$$

So, if integrals can be evaluated and simplified to explicit functions, then they should be worked out. If it is not possible, then answers in integral form are completely fine.

Classical problems leading to differential equations solvable by integration are problems dealing with velocity, acceleration and distance. You have surely seen these problems before in your calculus class.

Example 1.1.3: Suppose a car drives at a speed $e^{t/2}$ meters per second, where t is time in seconds. How far did the car get in 2 seconds (starting at $t = 0$)? How far in 10 seconds?

Solution: Let x denote the distance the car traveled. The equation is

$$x' = e^{t/2}.$$

We just integrate this equation to get that

$$x(t) = 2e^{t/2} + C.$$

We still need to figure out C . We know that when $t = 0$, then $x = 0$. That is, $x(0) = 0$. So

$$0 = x(0) = 2e^{0/2} + C = 2 + C.$$

Thus $C = -2$ and

$$x(t) = 2e^{t/2} - 2.$$

Now we just plug in to get where the car is at 2 and at 10 seconds. We obtain

$$x(2) = 2e^{2/2} - 2 \approx 3.44 \text{ meters}, \quad x(10) = 2e^{10/2} - 2 \approx 294 \text{ meters.}$$

□

Example 1.1.4: Suppose that the car accelerates at a rate of $t^2 \text{ m/s}^2$. At time $t = 0$ the car is at the 1 meter mark and is traveling at 10 m/s . Where is the car at time $t = 10$?

Solution: Well this is actually a second order problem. If x is the distance traveled, then x' is the velocity, and x'' is the acceleration. The initial value problem for this situation is

$$x'' = t^2, \quad x(0) = 1, \quad x'(0) = 10.$$

What if we say $x' = v$. Then we have the problem

$$v' = t^2, \quad v(0) = 10.$$

Once we solve for v , we can integrate and find x .

□

Exercise 1.1.1: Solve for v , and then solve for x . Find $x(10)$ to answer the question.

1.1.1 Exercises

Exercise 1.1.2: Solve $\frac{dy}{dx} = x^2 + x$ with $y(1) = 3$.

Exercise 1.1.3: Solve $\frac{dy}{dx} = \sin(5x)$ with $y(0) = 2$.

Exercise 1.1.4:* Solve $\frac{dy}{dx} = e^x + x$ with $y(0) = 10$.

Exercise 1.1.5: Solve $\frac{dy}{dx} = 2xe^{3x}$ with $y(0) = 1$.

Exercise 1.1.6: Solve $\frac{dx}{dt} = e^t \cos(2t) + t$ with $y(0) = 3$.

Exercise 1.1.7: Solve $\frac{dy}{dx} = \frac{1}{x^2+1} + 3e^{2x}$ with $y(0) = 2$.

Exercise 1.1.8: Solve $\frac{dy}{dx} = \frac{1}{x^2-1}$ for $y(0) = 0$.

Exercise 1.1.9 (harder): Solve $y'' = \sin x$ for $y(0) = 0$, $y'(0) = 2$.

Exercise 1.1.10: A spaceship is traveling at the speed $2t^2 + 1 \text{ km/s}$ (t is time in seconds). It is pointing directly away from earth and at time $t = 0$ it is 1000 kilometers from earth. How far from earth is it at one minute from time $t = 0$?

Exercise 1.1.11:* Sid is in a car traveling at speed $10t + 70$ miles per hour away from Las Vegas, where t is in hours. At $t = 0$, Sid is 10 miles away from Vegas. How far from Vegas is Sid 2 hours later?

Exercise 1.1.12: Solve $\frac{dx}{dt} = \sin(t^2) + t$, $x(0) = 20$. It is OK to leave your answer as a definite integral.

Exercise 1.1.13: Solve $\frac{dy}{dt} = e^{t^2} + \sin(t)$, $y(0) = 4$. The answer can be left as a definite integral.

Exercise 1.1.14: A dropped ball accelerates downwards at a constant rate 9.8 meters per second squared. Set up the differential equation for the height above ground h in meters. Then supposing $h(0) = 100$ meters, how long does it take for the ball to hit the ground.

Exercise 1.1.15:* The rate of change of the volume of a snowball that is melting is proportional to the surface area of the snowball. Suppose the snowball is perfectly spherical. The volume (in centimeters cubed) of a ball of radius r centimeters is $(4/3)\pi r^3$. The surface area is $4\pi r^2$. Set up the differential equation for how the radius r is changing. Then, suppose that at time $t = 0$ minutes, the radius is 10 centimeters. After 5 minutes, the radius is 8 centimeters. At what time t will the snowball be completely melted?

Exercise 1.1.16:* Find the general solution to $y''' = 0$. How many distinct constants do you need?

1.2 Slope fields

Attribution: [JL], §1.2.

Learning Objectives

After this section, you will be able to:

- Identify or sketch a slope field for a first order differential equation and
- Use the slope field to determine the trajectory of a solution to a differential equation.

As we said, the general first order equation we are studying looks like

$$y' = f(x, y).$$

A lot of the time, we cannot simply solve these kinds of equations explicitly, because our direct integration method only works when the equation is of the form $y' = f(x)$, which we could integrate directly. In these more complicated cases, it would be nice if we could at least figure out the shape and behavior of the solutions, or find approximate solutions.

1.2.1 Slope fields

Suppose that we have a solution to the equation $y' = f(x, y)$ with $y(x_0) = y_0$. What does the fact that this solves the differential equation tell us about the solution? It tells us that the derivative of the solution at this point will be $f(x_0, y_0)$. Graphically, the derivative gives the slope of the solution, so it means that the solution will pass through the point (x_0, y_0) and will have slope $f(x_0, y_0)$. For example, if $f(x, y) = xy$, then at point $(2, 1.5)$ we draw a short line of slope $xy = 2 \times 1.5 = 3$. So, if $y(x)$ is a solution and $y(2) = 1.5$, then the equation mandates that $y'(2) = 3$. See Figure 1.1.

To get an idea of how solutions behave, we draw such lines at lots of points in the plane, not just the point $(2, 1.5)$. We would ideally want to see the slope at every point, but that is just not possible. Usually we pick a grid of points fine enough so that it shows the behavior, but not too fine so that we can still recognize the individual lines. We call this picture the *slope field* of the equation. See Figure 1.2 on the next page for the slope field of the equation $y' = xy$. Usually in practice, one does not do this by hand, but has a computer do the drawing.

The idea of a slope field is that it tells us how the graph of the solution should be sloped, or should curve, if it passed through a given point. Having a wide variety of slopes plotted in our slope field gives an idea of how all of the solutions behave for a bunch of different initial

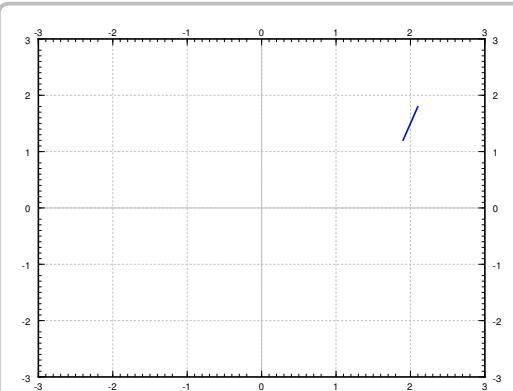


Figure 1.1: The slope $y' = xy$ at $(2, 1.5)$.

conditions. Which curve we want in particular, and where we should start the curve, depends on the initial condition.

Suppose we are given a specific initial condition $y(x_0) = y_0$. A solution, that is, the graph of the solution, would be a curve that follows the slopes we drew, starting from the point (x_0, y_0) . For a few sample solutions, see Figure 1.3. It is easy to roughly sketch (or at least imagine) possible solutions in the slope field, just from looking at the slope field itself. You simply sketch a line that roughly fits the little line segments and goes through your initial condition. The graph should “flow” along the little slopes that are on the slope field.

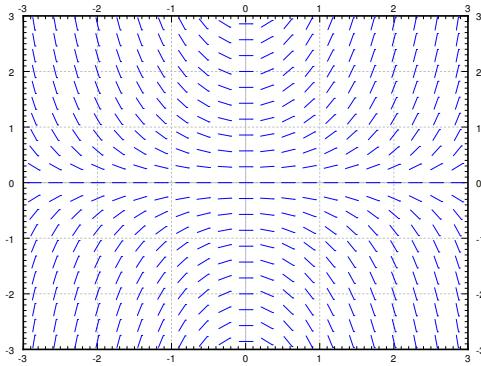


Figure 1.2: Slope field of $y' = xy$.

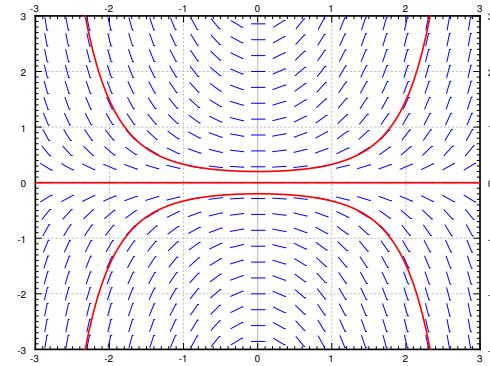


Figure 1.3: Slope field of $y' = xy$ with a graph of solutions satisfying $y(0) = 0.2$, $y(0) = 0$, and $y(0) = -0.2$.

By looking at the slope field we get a lot of information about the behavior of solutions without having to solve the equation. For example, in Figure 1.3 we see what the solutions do when the initial conditions are $y(0) > 0$, $y(0) = 0$ and $y(0) < 0$. A small change in the initial condition causes quite different behavior. We see this behavior just from the slope field and imagining what solutions ought to do.

We see a different behavior for the equation $y' = -y$. The slope field and a few solutions is in see Figure 1.4 on the following page. If we think of moving from left to right (perhaps x is time and time is usually increasing), then we see that no matter what $y(0)$ is, all solutions tend to zero as x tends to infinity. Again that behavior is clear from simply looking at the slope field itself.

1.2.2 Exercises

Exercise 1.2.1: Sketch slope field for $y' = e^{x-y}$. How do the solutions behave as x grows? Can you guess a particular solution by looking at the slope field?

Exercise 1.2.2:* Sketch the slope field of $y' = y^3$. Can you visually find the solution that satisfies $y(0) = 0$?

Exercise 1.2.3: Sketch slope field for $y' = x^2$.

Exercise 1.2.4: Sketch slope field for $y' = y^2$.

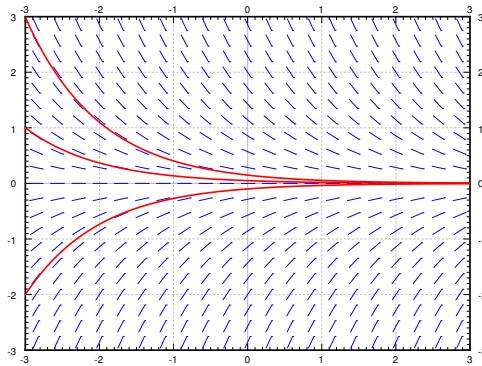
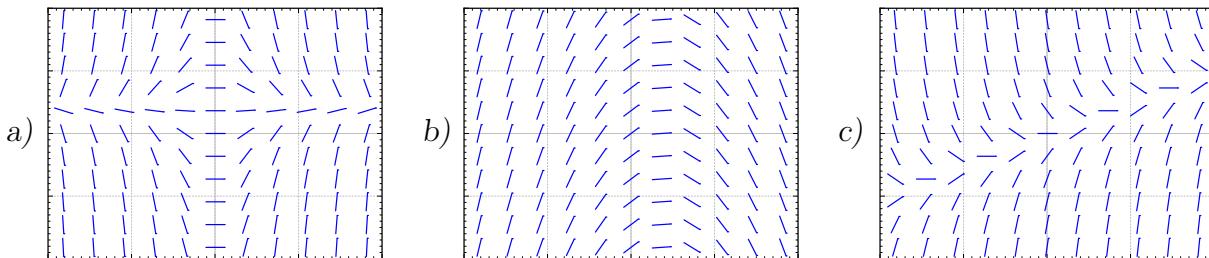


Figure 1.4: Slope field of $y' = -y$ with a graph of a few solutions.

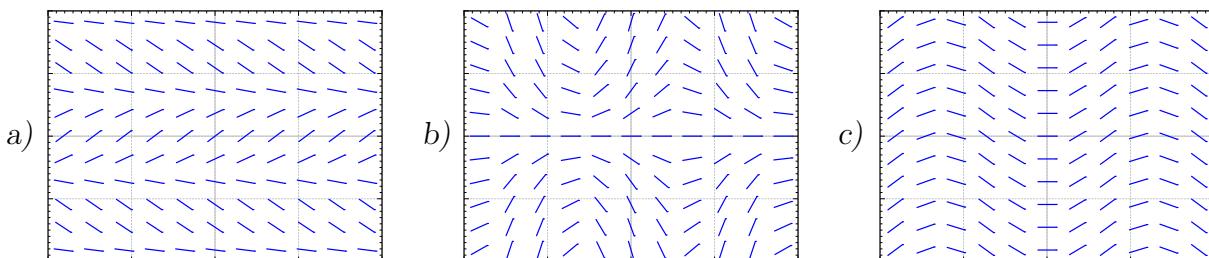
Exercise 1.2.5: For each of the following differential equations, sketch out a slope field on $-3 < x < 3$ and $-3 < y < 3$ and determine the overall behavior of the solutions to the equation as $t \rightarrow \infty$. If this fact depends on the value of the solution at $t = 0$, explain how it changes.

$$\text{a) } \frac{dy}{dx} = 3 - 2y \quad \text{b) } \frac{dy}{dx} = 1 + y \quad \text{c) } \frac{dy}{dx} = y - 1 \quad \text{d) } \frac{dy}{dx} = -2 - y$$

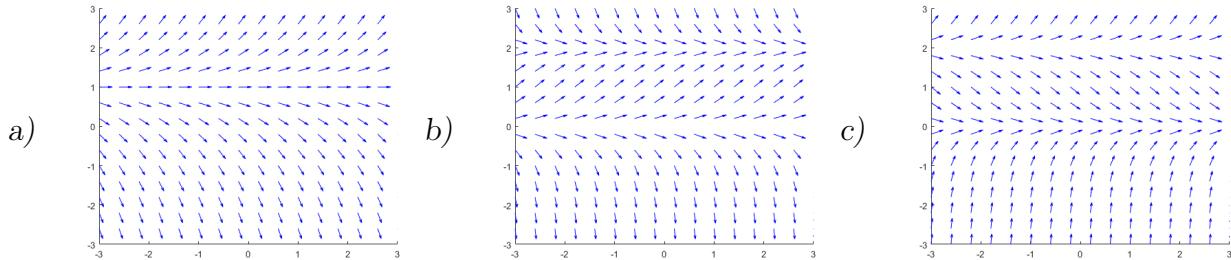
Exercise 1.2.6: Match equations $y' = 1 - x$, $y' = x - 2y$, $y' = x(1 - y)$ to slope fields. Justify.



Exercise 1.2.7:* Match equations $y' = \sin x$, $y' = \cos y$, $y' = y \cos(x)$ to slope fields. Justify.



Exercise 1.2.8: Match equations $y' = y(y - 2)$, $y' = y - 1$, $y' = y(2 - y)$ to slope fields. Justify.



Exercise 1.2.9 (challenging): Take $y' = f(x, y)$, $y(0) = 0$, where $f(x, y) > 1$ for all x and y . If the solution exists for all x , can you say what happens to $y(x)$ as x goes to positive infinity? Explain.

Exercise 1.2.10: Suppose $y' = f(x, y)$. What will the slope field look like, explain and sketch an example, if you know the following about $f(x, y)$:

- | | |
|---------------------------------------|--|
| a) f does not depend on y . | b) f does not depend on x . |
| c) $f(t, t) = 0$ for any number t . | d) $f(x, 0) = 0$ and $f(x, 1) = 1$ for all x . |

1.3 Separable equations

Attribution: [JL], §1.3.

Learning Objectives

After this section, you will be able to:

- Identify when a differential equation is separable,
- Find the general solution of a separable differential equation, and
- Solve initial value problems for separable differential equations.

As mentioned in the previous section, when a differential equation is of the form $y' = f(x)$, we can just integrate: $y = \int f(x) dx + C$. Unfortunately this method no longer works for the general form of the equation $y' = f(x, y)$. Integrating both sides yields

$$y = \int f(x, y) dx + C.$$

Notice the dependence on y in the integral. Since y is a function of x , this expression is really of the form

$$y = \int f(x, y(x)) dx + C$$

and without knowing what $y(x)$ is in advance (which we don't, because that's what we are trying to solve for) we can't compute this integral.

1.3.1 Separable equations

One particular type of differential equation that we can evaluate using a technique very similar to direct integration is separable equations.

Definition 1.3.1

We say a differential equation is *separable* if we can write it as

$$y' = f(x)g(y),$$

for some functions $f(x)$ and $g(y)$.

Let us write the equation in the Leibniz notation

$$\frac{dy}{dx} = f(x)g(y).$$

Then we rewrite the equation as

$$\frac{dy}{g(y)} = f(x) dx.$$

Both sides look like something we can integrate. We obtain

$$\int \frac{dy}{g(y)} = \int f(x) dx + C.$$

If we can find closed form expressions for these two integrals, we can, perhaps, solve for y .

Example 1.3.1: Solve the equation

$$y' = xy.$$

Solution: Note that $y = 0$ is a solution. We will remember that fact and assume $y \neq 0$ from now on, so that we can divide by y . Write the equation as $\frac{dy}{dx} = xy$. Then

$$\int \frac{dy}{y} = \int x dx + C.$$

We compute the antiderivatives to get

$$\ln|y| = \frac{x^2}{2} + C,$$

or

$$|y| = e^{\frac{x^2}{2} + C} = e^{\frac{x^2}{2}} e^C = D e^{\frac{x^2}{2}},$$

where $D > 0$ is some constant. Because $y = 0$ is also a solution and because of the absolute value we can write:

$$y = D e^{\frac{x^2}{2}},$$

for any number D (including zero or negative).

We check:

$$y' = D x e^{\frac{x^2}{2}} = x \left(D e^{\frac{x^2}{2}} \right) = xy.$$

Yay!

□

One particular case in which this method works very well is if the function $f(x, y)$ is only a function of y . With this, we can explicitly complete the solution to equations like

$$y' = ky,$$

reaching the solution $y(x) = e^{kx}$.

We should be a little bit more careful with this method. You may be worried that we integrated in two different variables. We seemingly did a different operation to each side. Let us work through this method more rigorously. Take

$$\frac{dy}{dx} = f(x)g(y).$$

We rewrite the equation as follows. Note that $y = y(x)$ is a function of x and so is $\frac{dy}{dx}$!

$$\frac{1}{g(y)} \frac{dy}{dx} = f(x).$$

We integrate both sides with respect to x :

$$\int \frac{1}{g(y)} \frac{dy}{dx} dx = \int f(x) dx + C.$$

We use the change of variables formula (substitution) on the left hand side:

$$\int \frac{1}{g(y)} dy = \int f(x) dx + C.$$

And we are done.

However, there are some special solutions to these problems as well that don't fit the same formula. Assume we have

$$\frac{dy}{dx} = f(x)g(y)$$

and we have a value y_0 such that $g(y_0) = 0$. Then, the function $y(x) = y_0$ is a solution, provided $f(x)$ is defined everywhere. (Plug this in and check!) This fills in the issue for having $\frac{1}{g(y)}$ in our integral expression, which is not defined when $g(y) = 0$. These are called *singular solutions*, and the next example will showcase one of them.

1.3.2 Implicit solutions

We sometimes get stuck even if we can do the integration. Consider the separable equation

$$y' = \frac{xy}{y^2 + 1}.$$

We separate variables,

$$\frac{y^2 + 1}{y} dy = \left(y + \frac{1}{y} \right) dy = x dx.$$

We integrate to get

$$\frac{y^2}{2} + \ln|y| = \frac{x^2}{2} + C,$$

or perhaps the easier looking expression (where $D = 2C$)

$$y^2 + 2\ln|y| = x^2 + D.$$

It is not easy to find the solution explicitly as it is hard to solve for y . We, therefore, leave the solution in this form and call it an *implicit solution*. It is still easy to check that an implicit solution satisfies the differential equation. In this case, we differentiate with respect to x , and remember that y is a function of x , to get

$$y' \left(2y + \frac{2}{y} \right) = 2x.$$

Multiply both sides by y and divide by $2(y^2 + 1)$ and you will get exactly the differential equation. We leave this computation to the reader.

If you have an implicit solution, and you want to compute values for y , you might have to be tricky. You might get multiple solutions y for each x , so you have to pick one. Sometimes you can graph x as a function of y , and then flip your paper. Sometimes you have to do more.

Computers are also good at some of these tricks. More advanced mathematical software usually has some way of plotting solutions to implicit equations, which makes these solutions just as good for visualizing or graphing as explicit solutions. For example, for $C = 0$ if you plot all the points (x, y) that are solutions to $y^2 + 2 \ln |y| = x^2$, you find the two curves in Figure 1.5. This is not quite a graph of a function. For each x there are two choices of y . To find a function you would have to pick one of these two curves. You pick the one that satisfies your initial condition if you have one. For example, the top curve satisfies the condition $y(1) = 1$. So for each C we really got two solutions. As you can see, computing values from an implicit solution can be somewhat tricky. But sometimes, an implicit solution is the best we can do.

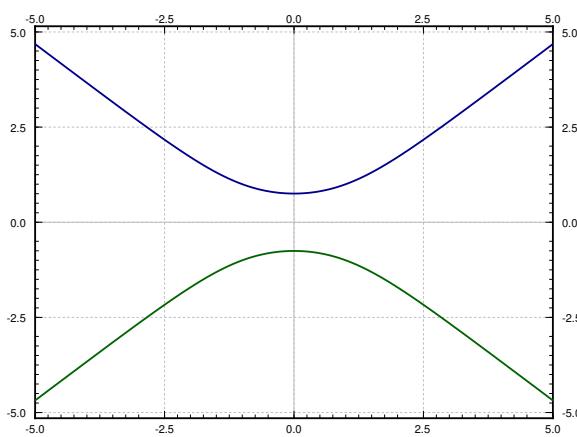


Figure 1.5: The implicit solution $y^2 + 2 \ln |y| = x^2$ to $y' = \frac{xy}{y^2+1}$.

The equation above also has the solution $y = 0$. Since our function

$$g(t) = \frac{y}{y^2 + 1}$$

this is zero at $y = 0$, and gives an additional solution to the problem. So the general solution is

$$y^2 + 2 \ln |y| = x^2 + C, \quad \text{and} \quad y = 0.$$

These outlying solutions such as $y = 0$ are sometimes called *singular solutions*.

1.3.3 Examples of separable equations

Example 1.3.2: Solve $x^2y' = 1 - x^2 + y^2 - x^2y^2$, $y(1) = 0$.

Solution: Factor the right-hand side

$$x^2y' = (1 - x^2)(1 + y^2).$$

Separate variables, integrate, and solve for y :

$$\begin{aligned}\frac{y'}{1+y^2} &= \frac{1-x^2}{x^2}, \\ \frac{y'}{1+y^2} &= \frac{1}{x^2} - 1, \\ \arctan(y) &= \frac{-1}{x} - x + C, \\ y &= \tan\left(\frac{-1}{x} - x + C\right).\end{aligned}$$

Solve for the initial condition, $0 = \tan(-2 + C)$ to get $C = 2$ (or $C = 2 + \pi$, or $C = 2 + 2\pi$, etc.). The particular solution we seek is, therefore,

$$y = \tan\left(\frac{-1}{x} - x + 2\right).$$

]

Example 1.3.3: Bob made a cup of coffee, and Bob likes to drink coffee only once reaches 60 degrees Celsius and will not burn him. Initially at time $t = 0$ minutes, Bob measured the temperature and the coffee was 89 degrees Celsius. One minute later, Bob measured the coffee again and it had 85 degrees. The temperature of the room (the ambient temperature) is 22 degrees. When should Bob start drinking?

Solution: Let T be the temperature of the coffee in degrees Celsius, and let A be the ambient (room) temperature, also in degrees Celsius. Newton's law of cooling states that the rate at which the temperature of the coffee is changing is proportional to the difference between the ambient temperature and the temperature of the coffee. That is,

$$\frac{dT}{dt} = k(A - T),$$

for some constant k . For our setup $A = 22$, $T(0) = 89$, $T(1) = 85$. We separate variables and integrate (let C and D denote arbitrary constants):

$$\begin{aligned}\frac{1}{T-A} \frac{dT}{dt} &= -k, \\ \ln(T-A) &= -kt + C, \quad (\text{note that } T-A > 0) \\ T-A &= D e^{-kt}, \\ T &= A + D e^{-kt}.\end{aligned}$$

That is, $T = 22 + D e^{-kt}$. We plug in the first condition: $89 = T(0) = 22 + D$, and hence $D = 67$. So $T = 22 + 67 e^{-kt}$. The second condition says $85 = T(1) = 22 + 67 e^{-k}$. Solving for k we get $k = -\ln \frac{85-22}{67} \approx 0.0616$. Now we solve for the time t that gives us a temperature of 60 degrees. Namely, we solve

$$60 = 22 + 67 e^{-0.0616t}$$

to get $t = -\frac{\ln \frac{60-22}{67}}{0.0616} \approx 9.21$ minutes. So Bob can begin to drink the coffee at just over 9 minutes from the time Bob made it. That is probably about the amount of time it took us to calculate how long it would take. See [Figure 1.6](#) on the facing page.

]

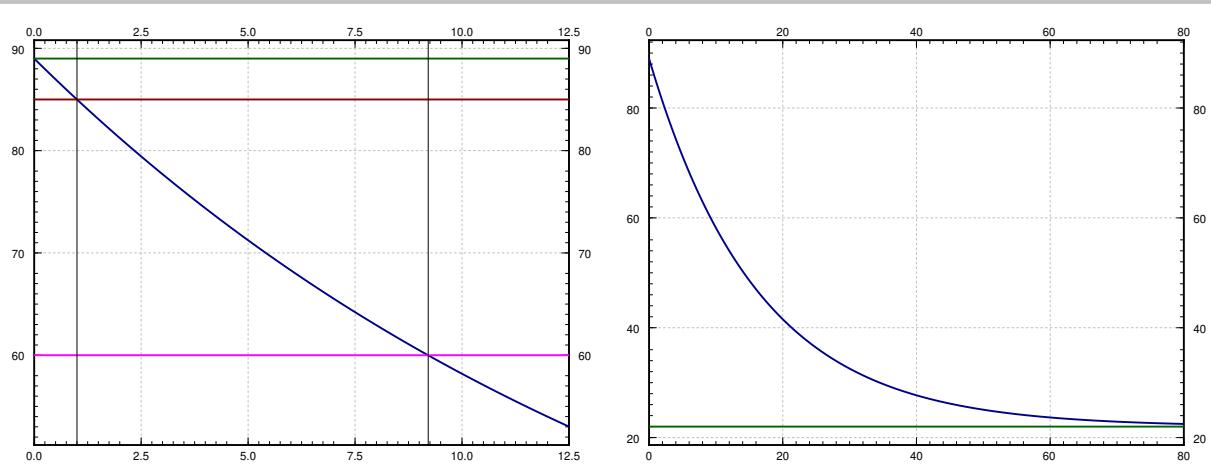


Figure 1.6: Graphs of the coffee temperature function $T(t)$. On the left, horizontal lines are drawn at temperatures 60, 85, and 89. Vertical lines are drawn at $t = 1$ and $t = 9.21$. Notice that the temperature of the coffee hits 85 at $t = 1$, and 60 at $t \approx 9.21$. On the right, the graph is over a longer period of time, with a horizontal line at the ambient temperature 22.

Example 1.3.4: Find the general solution to $y' = \frac{-xy^2}{3}$ (including singular solutions).

Solution: First note that $y = 0$ is a solution (a singular solution). Now assume that $y \neq 0$.

$$\begin{aligned} \frac{-3}{y^2} y' &= x, \\ \frac{3}{y} &= \frac{x^2}{2} + C, \\ y &= \frac{3}{x^2/2 + C} = \frac{6}{x^2 + 2C}. \end{aligned}$$

So the general solution is,

$$y = \frac{6}{x^2 + 2C}, \quad \text{and} \quad y = 0.$$

□

1.3.4 Exercises

Exercise 1.3.1: Solve $y' = y^3$ for $y(0) = 1$.

Exercise 1.3.2:* Solve $x' = \frac{1}{x^2}$, $x(1) = 1$.

Exercise 1.3.3 (little harder): Solve $y' = (y - 1)(y + 1)$ for $y(0) = 3$.

Exercise 1.3.4:* Solve $x' = \frac{1}{\cos(x)}$, $x(0) = \frac{\pi}{2}$.

Exercise 1.3.5: Solve $\frac{dy}{dx} = \frac{1}{y+1}$ for $y(0) = 0$.

Exercise 1.3.6: Solve $y' = x/y$.

Exercise 1.3.7: Solve $y' = x^2y$.

Exercise 1.3.8:* Consider the differential equation

$$\frac{dy}{dx} = \frac{2x}{y}$$

- a) Find the general solution as an implicit function.
- b) Find the solution to this differential equation as an explicit function with $y(1) = 4$.
- c) Find the solution to this differential equation as an explicit function with $y(0) = -2$.

Exercise 1.3.9:* Solve $y' = y^n$, $y(0) = 1$, where n is a positive integer. Hint: You have to consider different cases.

Exercise 1.3.10: Solve $\frac{dx}{dt} = (x^2 - 1)t$, for $x(0) = 0$.

Exercise 1.3.11: Solve $\frac{dx}{dt} = x \sin(t)$, for $x(0) = 1$.

Exercise 1.3.12:* Solve $y' = 2xy$.

Exercise 1.3.13: Solve $\frac{dy}{dx} = xy + x + y + 1$. Hint: Factor the right-hand side.

Exercise 1.3.14:* Solve $x' = 3xt^2 - 3t^2$, $x(0) = 2$.

Exercise 1.3.15: Find the general solution of $y' = e^x$, and then $y' = e^y$.

Exercise 1.3.16: Solve $xy' = y + 2x^2y$, where $y(1) = 1$.

Exercise 1.3.17:* Find an implicit solution for $x' = \frac{1}{3x^2+1}$, $x(0) = 1$.

Exercise 1.3.18: Solve $\frac{dy}{dx} = \frac{y^2 + 1}{x^2 + 1}$, for $y(0) = 1$.

Exercise 1.3.19: Find an implicit solution for $\frac{dy}{dx} = \frac{x^2 + 1}{y^2 + 1}$, for $y(0) = 1$.

Exercise 1.3.20:* Find an implicit solution to $y' = \frac{\sin(x)}{\cos(y)}$.

Exercise 1.3.21: Find an implicit solution for $xy' = \frac{x^2+1}{y^2-1}$ with $y(3) = 2$.

Exercise 1.3.22: Find an explicit solution for $y' = xe^{-y}$, $y(0) = 1$.

Exercise 1.3.23:* Find an explicit solution to $xy' = y^2$, $y(1) = 1$.

Exercise 1.3.24: Find an explicit solution for $xy' = e^{-y}$, for $y(1) = 1$.

Exercise 1.3.25: Find an explicit solution for $y' = y^2(x^4 + 1)$ with $y(1) = 2$.

Exercise 1.3.26: Find an explicit solution for $y' = \frac{\cos(x)+1}{y}$ with $y(0) = 4$.

Exercise 1.3.27: Find an explicit solution for $y' = ye^{-x^2}$, $y(0) = 1$. It is alright to leave a definite integral in your answer.

Exercise 1.3.28: Suppose a cup of coffee is at 100 degrees Celsius at time $t = 0$, it is at 70 degrees at $t = 10$ minutes, and it is at 50 degrees at $t = 20$ minutes. Compute the ambient temperature.

Exercise 1.3.29:* Take [Example 1.3.3](#) with the same numbers: 89 degrees at $t = 0$, 85 degrees at $t = 1$, and ambient temperature of 22 degrees. Suppose these temperatures were measured with precision of ± 0.5 degrees. Given this imprecision, the time it takes the coffee to cool to (exactly) 60 degrees is also only known in a certain range. Find this range. Hint: Think about what kind of error makes the cooling time longer and what shorter.

Exercise 1.3.30:* A population x of rabbits on an island is modeled by $x' = x - (1/1000)x^2$, where the independent variable is time in months. At time $t = 0$, there are 40 rabbits on the island.

- a) Find the solution to the equation with the initial condition.
- b) How many rabbits are on the island in 1 month, 5 months, 10 months, 15 months (round to the nearest integer).

1.4 Linear equations and the integrating factor

Attribution: [JL], §1.4.

Learning Objectives

After this section, you will be able to:

- Identify a linear first-order differential equation and write a first-order linear equation in standard form,
- Solve initial value problems for first-order linear differential equations by integrating factors, and
- Write solutions to first-order linear initial value problems in integral form if needed.

One of the most important types of equations we will learn how to solve are the so-called *linear equations*. In fact, the majority of the course is about linear equations. In this section we focus on the *first order linear equation*.

Definition 1.4.1

A first order equation is *linear* if we can put it into the form:

$$y' + p(x)y = f(x). \quad (1.3)$$

The word “linear” means linear in y and y' ; no higher powers nor functions of y or y' appear. The dependence on x can be more complicated.

Solutions of linear equations have nice properties. For example, the solution exists wherever $p(x)$ and $f(x)$ are defined, and has the same regularity (read: it is just as nice). We'll see this in detail in § 1.5. But most importantly for us right now, there is a method for solving linear first order equations. In § 1.1, we saw that we could easily solve equations of the form

$$\frac{dy}{dx} = f(x)$$

because we could directly integrate both sides of the equation, since the left hand side was the derivative of something (in this case, y) and the right side was only a function of x . We want to do the same here, but the something on the left will not just be y .

The trick is to rewrite the left-hand side of (1.3) as a derivative of a product of y with another function. Let $r(x)$ be this other function, and we can compute, by the product rule, that

$$\frac{d}{dx} [r(x)y] = r(x)y' + r'(x)y.$$

Now, if we multiply (1.3) by the function $r(x)$ on both sides, we get

$$r(x)y' + p(x)r(x)y = f(x)r(x)$$

and the first term on the left here matches the first term from our product rule derivative. To make the second terms match up as well, we need that

$$r'(x) = p(x)r(x).$$

This equation is separable! We can solve for the $r(x)$ here by separating variables to get that

$$\frac{dr}{r} = p(x) dx$$

so that

$$\ln |r| = \int p(x) dx$$

or

$$r(x) = e^{\int p(x) dx}.$$

With this choice of $r(x)$, we get that

$$r(x)y' + r(x)p(x)y = \frac{d}{dx} [r(x)y],$$

so that if we multiply (1.3) by $r(x)$, we obtain $r(x)y' + r(x)p(x)y$ on the left-hand side, which we can simplify using our product rule derivative above to obtain

$$\frac{d}{dx} [r(x)y] = r(x)f(x).$$

Now we integrate both sides. The right-hand side does not depend on y and the left-hand side is written as a derivative of a function. Afterwards, we solve for y . The function $r(x)$ is called the *integrating factor* and the method is called the *integrating factor method*.

This method works for any first order linear equation, no matter what $p(x)$ and $f(x)$ are. In general, we can compute:

$$\begin{aligned} y' + p(x)y &= f(x), \\ e^{\int p(x) dx} y' + e^{\int p(x) dx} p(x)y &= e^{\int p(x) dx} f(x), \\ \frac{d}{dx} \left[e^{\int p(x) dx} y \right] &= e^{\int p(x) dx} f(x), \\ e^{\int p(x) dx} y &= \int e^{\int p(x) dx} f(x) dx + C, \\ y &= e^{-\int p(x) dx} \left(\int e^{\int p(x) dx} f(x) dx + C \right). \end{aligned}$$

Of course, to get a closed form formula for y , we need to be able to find a closed form formula for the integrals appearing above.

Example 1.4.1: Solve

$$y' + 2xy = e^{x-x^2}, \quad y(0) = -1.$$

Solution: First note that $p(x) = 2x$ and $f(x) = e^{x-x^2}$. The integrating factor is $r(x) = e^{\int p(x) dx} = e^{x^2}$. We multiply both sides of the equation by $r(x)$ to get

$$\begin{aligned} e^{x^2} y' + 2xe^{x^2} y &= e^{x-x^2} e^{x^2}, \\ \frac{d}{dx} [e^{x^2} y] &= e^x. \end{aligned}$$

We integrate

$$\begin{aligned} e^{x^2} y &= e^x + C, \\ y &= e^{x-x^2} + Ce^{-x^2}. \end{aligned}$$

Next, we solve for the initial condition $-1 = y(0) = 1 + C$, so $C = -2$. The solution is

$$y = e^{x-x^2} - 2e^{-x^2}.$$

□

Note that we do not care which antiderivative we take when computing $e^{\int p(x) dx}$. You can always add a constant of integration, but those constants will not matter in the end.

Exercise 1.4.1: Try it! Add a constant of integration to the integral in the integrating factor and show that the solution you get in the end is the same as what we got above.

Advice: Do not try to remember the formula itself, that is way too hard. It is easier to remember the process and repeat it.

Since we cannot always evaluate the integrals in closed form, it is useful to know how to write the solution in definite integral form. A definite integral is something that you can plug into a computer or a calculator. Suppose we are given

$$y' + p(x)y = f(x), \quad y(x_0) = y_0.$$

Look at the solution and write the integrals as definite integrals.

$$y(x) = e^{-\int_{x_0}^x p(s) ds} \left(\int_{x_0}^x e^{\int_{x_0}^t p(s) ds} f(t) dt + y_0 \right). \quad (1.4)$$

You should be careful to properly use dummy variables here. If you now plug such a formula into a computer or a calculator, it will be happy to give you numerical answers.

Exercise 1.4.2: Check that $y(x_0) = y_0$ in formula (1.4).

Exercise 1.4.3: Write the solution of the following problem as a definite integral, but try to simplify as far as you can. You will not be able to find the solution in closed form.

$$y' + y = e^{x^2-x}, \quad y(0) = 10.$$

1.4.1 Exercises

In the exercises, feel free to leave answer as a definite integral if a closed form solution cannot be found. If you can find a closed form solution, you should give that.

Exercise 1.4.4: Solve $y' + xy = x$.

Exercise 1.4.5: Solve $y' + 6y = e^x$.

Exercise 1.4.6: Solve $y' + 4y = x^2e^{-4x}$.

Exercise 1.4.7: Solve $y' - 3y = xe^x$.

Exercise 1.4.8: Solve $y' + 3y = e^{4x} - e^{-2x}$ with $y(0) = -3$.

Exercise 1.4.9: Solve $y' - 2y = x + 4$.

Exercise 1.4.10: Solve $xy' + 4y = x^2 - \frac{1}{x^2}$.

Exercise 1.4.11: Solve $xy' - 3y = x - 2$ with $y(1) = 3$.

Exercise 1.4.12: Solve $y' - 4y = \cos(3t)$.

Exercise 1.4.13:* Solve $y' + 3x^2y = x^2$.

Exercise 1.4.14: Solve $y' + 3x^2y = \sin(x)e^{-x^3}$, with $y(0) = 1$.

Exercise 1.4.15: Solve $y' + \cos(x)y = \cos(x)$.

Exercise 1.4.16: Solve $\frac{1}{x^2+1}y' + xy = 3$, with $y(0) = 0$.

Exercise 1.4.17:* Solve $y' + 2\sin(2x)y = 2\sin(2x)$, $y(\pi/2) = 3$.

Exercise 1.4.18: Consider the initial value problem

$$5y' - 3y = e^{-2t} \quad y(0) = a$$

for an undetermined value a . Solve the problem and determine the dependence on the the value of a . How does the value of the solution as $t \rightarrow \infty$ depend on the value of a ?

Exercise 1.4.19: Find an expression for the general solution to $y' + 3y = \sin(t^2)$ with $y(0) = 2$. Simplify your answer as much as possible.

1.5 Existence and Uniqueness of Solutions

Attribution: [JL], §1.2.

Learning Objectives

After this section, you will be able to:

- Understand the terms existence and uniqueness as they apply to differential equations and
- Find the maximum guaranteed interval of existence for the solution to an initial value problem.

If we take the differential equation

$$y' = f(x, y) \quad y(x_0) = y_0,$$

there are two main questions we want to answer about this equation.

- Does a solution exist to the differential equation?
- Is there only one solution to the differential equation?

These are more commonly referred to as (a) existence of the solution and (b) uniqueness of the solution. These are especially crucial for equations that we are using to model a physical situation. For physical situations, the solution definitely exists (because the system does something and continues to exist) and the solution is unique, because a given system will always do the same thing given the same setup. Since we know that physical systems obey these properties, the equations we use to model them should have these properties as well. These properties do not necessarily hold for all differential equations, as shown in the examples below.

Example 1.5.1: Attempt to solve:

$$y' = \frac{1}{x}, \quad y(0) = 0.$$

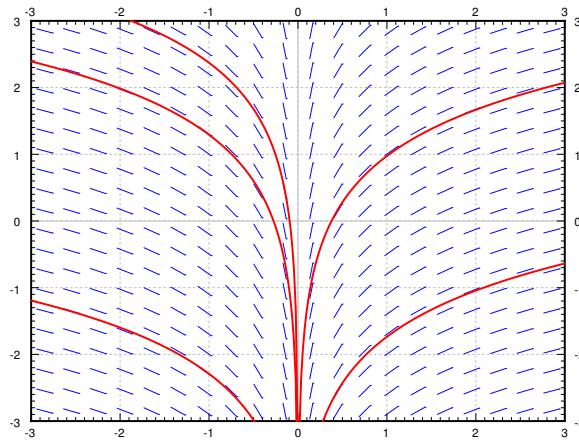
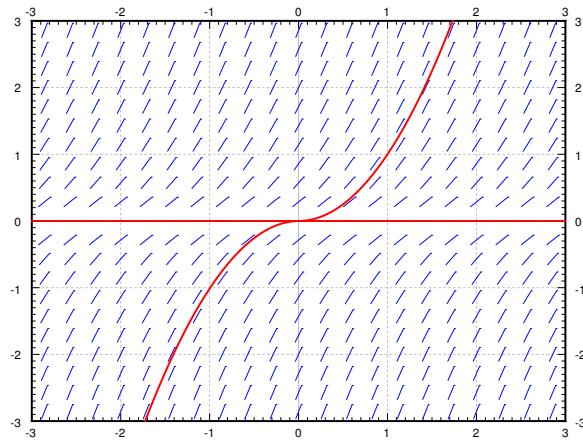
Integrate to find the general solution $y = \ln|x| + C$. The solution does not exist at $x = 0$. See [Figure 1.7](#) on the next page. The equation may have been written as the seemingly harmless $xy' = 1$.

Example 1.5.2: Solve:

$$y' = 2\sqrt{|y|}, \quad y(0) = 0.$$

See [Figure 1.8](#) on the facing page. Note that $y = 0$ is a solution. But another solution is the function

$$y(x) = \begin{cases} x^2 & \text{if } x \geq 0, \\ -x^2 & \text{if } x < 0. \end{cases}$$

Figure 1.7: Slope field of $y' = 1/x$.Figure 1.8: Slope field of $y' = 2\sqrt{|y|}$ with two solutions satisfying $y(0) = 0$.

Thankfully, these properties do apply to most equations, and we have fairly straightforward criteria that can be used to determine if these properties exist for a given differential equation. For a first-order linear differential equation, the theorem is fairly straight-forward.

Theorem 1.5.1

Assume that we have the first-order linear differential equation given by

$$y' + p(x)y = g(x).$$

If $p(x)$ and $g(x)$ are continuous functions on an interval I that contains a point x_0 , then for any y -value y_0 , the initial value problem

$$y' + p(x)y = g(x) \quad y(x_0) = y_0$$

has a unique solution. This solution exists and is unique on the entire interval I .

The idea and proof of this theorem comes from the fact that we have an explicit method for solving these equations no matter what p and g are. We can always find an integrating factor for the equation, convert the left-hand side into a product rule term, take a definite integral of both sides, and then solve for y . Since we have this explicit formula, the solution will exist and be defined on the entire interval where the functions p and g are continuous. This also means that we can answer questions about where and for what values of x the solution to a differential equation exists.

Example 1.5.3: Consider the differential equation

$$(x - 1)y' + \frac{1}{x - 5}y = e^x$$

What do the existence and uniqueness theorems say about the solution to this differential

equation with the initial condition $y(2) = 6$? What about the solution with initial condition $y(-3) = 1$?

Solution: To apply the existence and uniqueness theorem, we need to get the y' term by itself. This results in the differential equation

$$y' + \frac{1}{(x-1)(x-5)}y = \frac{e^x}{x-1}.$$

In order to figure out where this solution exists and is unique, we need to determine where the coefficient functions $p(x)$ and $g(x)$ are continuous. The only two points that we have discontinuities are at $x = 1$ and $x = 5$. Therefore, if we have the initial condition $y(2) = 6$, we start at the x value of 2. Because this equation is linear, it will exist everywhere that these two functions are both continuous containing the point $x = 2$, and since the only discontinuities are at 1 and 5, we know that they are both continuous on $(1, 5)$. This means that we can take $(1, 5)$ as the interval I in the theorem, and know that this solution will exist and be unique on the interval $(1, 5)$.

For the other initial condition, $y(-3) = 1$, we now want an interval where these functions are continuous that contains -3 . Again, we only have to avoid $x = 1$ and $x = 5$, so we can take the interval $(-\infty, 1)$ as the interval I in the theorem, and so we know the solution with this initial condition will exist and be unique on $(-\infty, 1)$. \square

For non-linear equations, we don't have an explicit method of getting a solution that works for all equations. This means that we can't fall back on this formula to guarantee existence or uniqueness of solutions. For this reason, we expect to get a result that is not as strong for non-linear equations. Thankfully, we do still get a result, which is known as Picard's theorem*.

Theorem 1.5.2 (Picard's theorem on existence and uniqueness)

If $f(x, y)$ is continuous (as a function of two variables) and $\frac{\partial f}{\partial y}$ exists and is continuous near some (x_0, y_0) , then a solution to

$$y' = f(x, y), \quad y(x_0) = y_0,$$

exists (at least for some small interval of x 's) and is unique.

The main fact that is “not as strong” about this result is the interval that we get from the theorem. For the linear theorem, we got existence and uniqueness on the entire interval I where p and g are continuous. For the non-linear theorem, we only get existence on *some* interval around the point x_0 . Even if $f(x, y)$ and $\frac{\partial f}{\partial y}$ are really nice functions that are continuous everywhere, we can still only guarantee existence on a small interval (that can depend on the initial condition) around the point x_0 .

Example 1.5.4: For some constant A , solve:

$$y' = y^2, \quad y(0) = A.$$

*Named after the French mathematician [Charles Émile Picard](#) (1856–1941)

Solution: We know how to solve this equation. First assume that $A \neq 0$, so y is not equal to zero at least for some x near 0. So $x' = 1/y^2$, so $x = -1/y + C$, so $y = \frac{1}{C-x}$. If $y(0) = A$, then $C = 1/A$ so

$$y = \frac{1}{1/A - x}.$$

If $A = 0$, then $y = 0$ is a solution.

For example, when $A = 1$ the solution is

$$y = \frac{1}{1-x}$$

which goes to infinity, and so “blows up”, at $x = 1$. This solution here exists only on the interval $(-\infty, 1)$, and hence, the solution does not exist for all x even if the equation is nice everywhere. The equation $y' = y^2$ certainly looks nice.

However, this fact does not contradict our existence and uniqueness theorem for non-linear equations. The theorem only guarantees that the solution to

$$y' = \frac{1}{y^2}$$

exists and is unique on *some* interval containing 0. It does not guarantee that the solution exists everywhere that $\frac{1}{y^2}$ and its derivative are continuous, only that at each point where this happens, the solution will exist for some interval around that point. The interval $(-\infty, 1)$ is “some interval containing 0”, so the theorem still applies and holds here. \square

The other main conclusion that we can draw from these theorems is the fact that two different solution curves to a first-order differential equation can not cross, provided the existence and uniqueness theorems hold. If y_1 and y_2 are two different solutions to $y' = f(x, y)$ and the solution curves for $y_1(x)$ and $y_2(x)$ cross, then this means that for some particular value of x_0 and y_0 , we have that

$$y_1(x_0) = y_0 \quad y_2(x_0) = y_0.$$

If we pick x_0 as a starting point, then the fact that the existence and uniqueness theorems hold imply that, at least for some interval around x_0 , there is exactly one solution to

$$y' = f(x, y) \quad y(x_0) = y_0.$$

However, both y_1 and y_2 satisfy these two properties. Therefore, y_1 and y_2 must be the same, which doesn’t make sense because we assumed they were different. So it is impossible for two different solution curves to cross, provided the existence and uniqueness theorem holds.

This fact is useful for analyzing differential equations in general, but will be particularly useful in § 1.7 in dealing with autonomous equations, where we can use simple solutions to provide boundaries over which other solutions can not cross. This fact will come up again in Chapters 4 and 5 in sketching trajectories for these solutions as well.

1.5.1 Exercises

Exercise 1.5.1: Is it possible to solve the equation $y' = \frac{xy}{\cos x}$ for $y(0) = 1$? Justify.

Exercise 1.5.2: Is it possible to solve the equation $y' = y\sqrt{|x|}$ for $y(0) = 0$? Is the solution unique? Justify.

Exercise 1.5.3: Consider the differential equation $y' + \frac{1}{t-2}y = \frac{1}{t+3}$.

- a) Is this equation linear or non-linear?
- b) What is the maximum guaranteed interval of existence for the solution to this equation with initial condition $y(0) = 3$?
- c) What if we start with the initial condition $y(4) = 0$?

Exercise 1.5.4: Consider the differential equation $y' + \frac{1}{t+2}y = \frac{\ln(|t|)}{t-4}$.

- a) Is this equation linear or non-linear?
- b) What is the maximum guaranteed interval of existence for the solution to this equation with initial condition $y(-3) = 1$?
- c) What if we start with the initial condition $y(2) = 5$?
- d) What happens if we want to start with $y(4) = 3$?

Exercise 1.5.5: Consider the differential equation $(t+3)y' + t^2y = \frac{1}{t-2}$.

- a) Is this equation linear or non-linear?
- b) What is the maximum guaranteed interval of existence for the solution to this equation with initial condition $y(-2) = 1$?
- c) What if we start with the initial condition $y(-4) = 5$?
- d) What happens if we want to start with $y(4) = 2$?

Exercise 1.5.6: Consider the differential equation $y' = y^2$.

- a) Is this equation linear or non-linear?
- b) What is the maximum guaranteed interval of existence for the solution to this equation with initial condition $y(0) = 1$?
- c) Find the solution to this differential equation with $y(0) = 1$. Over what values of x does this solution exist?
- d) Find the solution to this differential equation with $y(0) = 4$. Over what values of x does this solution exist?
- e) Find the solution to this differential equation with $y(0) = -2$. Over what values of x does this solution exist?
- f) Do any of these violate your answer in (b)?

Exercise 1.5.7: Consider the differential equation $y' = y^2 + 4$.

- a) Is this equation linear or non-linear?
- b) What is the maximum guaranteed interval of existence for the solution to this equation with initial condition $y(0) = 0$?
- c) Find the solution to this differential equation with $y(0) = 0$. Over what values of x does this solution exist?

Exercise 1.5.8: Consider the differential equation $y' = x(y + 1)^2$.

- a) Is this equation linear or non-linear?
- b) If we set $f(x, y) = x(y + 1)^2$, for what values of x and y are f and $\frac{\partial f}{\partial y}$ continuous?
- c) What is the maximum guaranteed interval of existence for the solution to this equation with initial condition $y(0) = 1$?
- d) Find the solution to this differential equation with $y(0) = 1$. Over what values of x does this solution exist?

Exercise 1.5.9 (challenging): Take $(y - x)y' = 0$, $y(0) = 0$.

- a) Find two distinct solutions.
- b) Explain why this does not violate Picard's theorem.

Exercise 1.5.10: Find a solution to $y' = |y|$, $y(0) = 0$. Does Picard's theorem apply?

Exercise 1.5.11: Take an equation $y' = (y - 2x)g(x, y) + 2$ for some function $g(x, y)$. Can you solve the problem for the initial condition $y(0) = 0$, and if so what is the solution?

Exercise 1.5.12: Consider the differential equation $y' = e^x(y - 2)$.

- a) Verify that $y = 2$ is a solution to this differential equation.
- b) Assume that we look for the solution with $y(0) = 0$. Is it possible that $y(x) = 3$ for some later time x ? Why or why not?
- c) Based on this, what do we know about the solution with $y(0) = 5$?

Exercise 1.5.13 (challenging): Suppose $y' = f(x, y)$ is such that $f(x, 1) = 0$ for every x , f is continuous and $\frac{\partial f}{\partial y}$ exists and is continuous for every x and y .

- a) Guess a solution given the initial condition $y(0) = 1$.
- b) Can graphs of two solutions of the equation for different initial conditions ever intersect?
- c) Given $y(0) = 0$, what can you say about the solution. In particular, can $y(x) > 1$ for any x ? Can $y(x) = 1$ for any x ? Why or why not?

Exercise 1.5.14:* Is it possible to solve $y' = xy$ for $y(0) = 0$? Is the solution unique?

Exercise 1.5.15: Is it possible to solve $y' = \frac{x}{x^2-1}$ for $y(1) = 0$?

Exercise 1.5.16 (tricky):* Suppose

$$f(y) = \begin{cases} 0 & \text{if } y > 0, \\ 1 & \text{if } y \leq 0. \end{cases}$$

Does $y' = f(y)$, $y(0) = 0$ have a continuously differentiable solution? Does Picard apply? Why, or why not?

Exercise 1.5.17:* Consider an equation of the form $y' = f(x)$ for some continuous function f , and an initial condition $y(x_0) = y_0$. Does a solution exist for all x ? Why or why not?

1.6 Numerical methods: Euler's method

Attribution: [JL], §1.7.

Learning Objectives

After this section, you will be able to:

- Use Euler's method to numerically approximate solutions to first order differential equations,
- Compute the error in a numerical method using the true solution, and
- Compare a variety of numerical methods, including built-in Matlab methods.

Unless $f(x, y)$ is of a special form, it is generally very hard if not impossible to get a nice formula for the solution of the problem

$$y' = f(x, y), \quad y(x_0) = y_0.$$

If the equation can be solved in closed form, we should do that. But what if we have an equation that cannot be solved in closed form? What if we want to find the value of the solution at some particular x ? Or perhaps we want to produce a graph of the solution to inspect the behavior. In this section we will learn about the basics of numerical approximation of solutions.

The simplest method for approximating a solution is *Euler's method*^{*}. It works as follows: Take x_0 and compute the slope $k = f(x_0, y_0)$. The slope is the change in y per unit change in x . Follow the line for an interval of length h on the x -axis. Hence if $y = y_0$ at x_0 , then we say that y_1 (the approximate value of y at $x_1 = x_0 + h$) is $y_1 = y_0 + hk$. Rinse, repeat! Let $k = f(x_1, y_1)$, and then compute $x_2 = x_1 + h$, and $y_2 = y_1 + hk$. Now compute x_3 and y_3 using x_2 and y_2 , etc. Consider the equation $y' = y^2/3$, $y(0) = 1$, and $h = 1$. Then $x_0 = 0$ and $y_0 = 1$. We compute

$$\begin{aligned} x_1 &= x_0 + h = 0 + 1 = 1, & y_1 &= y_0 + h f(x_0, y_0) = 1 + 1 \cdot 1/3 = 4/3 \approx 1.333, \\ x_2 &= x_1 + h = 1 + 1 = 2, & y_2 &= y_1 + h f(x_1, y_1) = 4/3 + 1 \cdot \frac{(4/3)^2}{3} = 52/27 \approx 1.926. \end{aligned}$$

We then draw an approximate graph of the solution by connecting the points (x_0, y_0) , (x_1, y_1) , (x_2, y_2) , For the first two steps of the method see Figure 1.9 on the next page.

More abstractly, for any $i = 0, 1, 2, 3, \dots$, we compute

$$x_{i+1} = x_i + h, \quad y_{i+1} = y_i + h f(x_i, y_i).$$

This can be worked out by hand for a few steps, but the formulas here lend themselves very well to being coded into a looping structure for more involved processes. The line segments we get are an approximate graph of the solution. Generally it is not exactly the solution. See Figure 1.10 on the following page for the plot of the real solution and the approximation.

^{*}Named after the Swiss mathematician Leonhard Paul Euler (1707–1783). The correct pronunciation of the name sounds more like “oiler.”

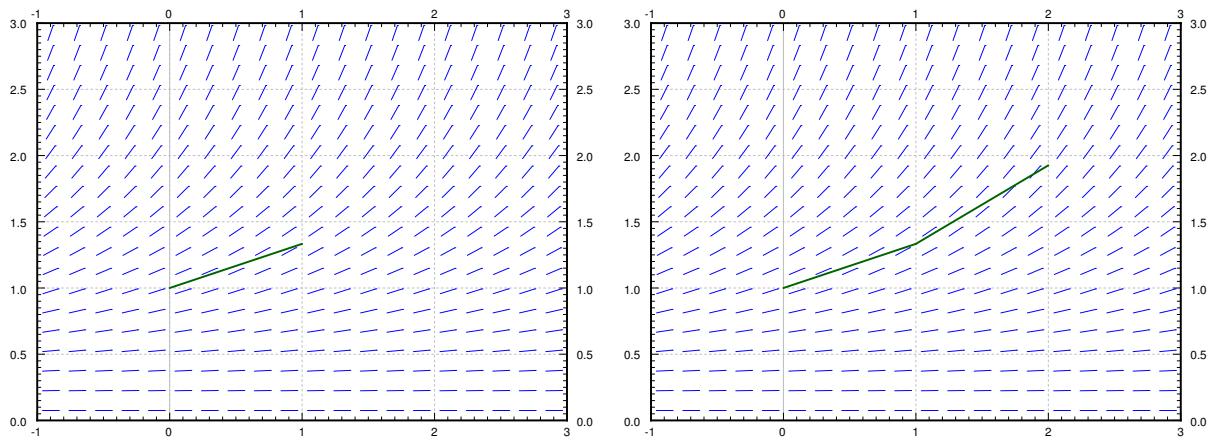


Figure 1.9: First two steps of Euler's method with $h = 1$ for the equation $y' = \frac{y^2}{3}$ with initial conditions $y(0) = 1$.

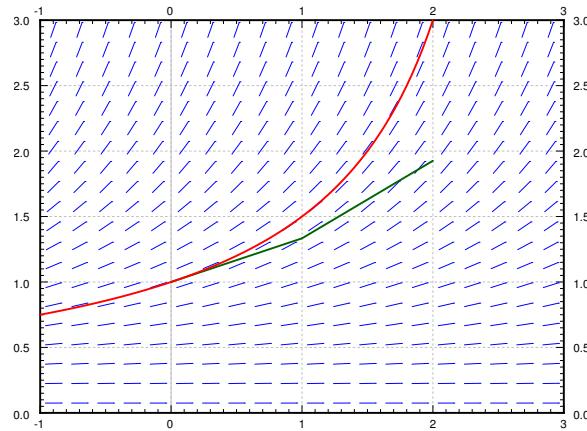


Figure 1.10: Two steps of Euler's method (step size 1) and the exact solution for the equation $y' = \frac{y^2}{3}$ with initial conditions $y(0) = 1$.

We continue with the equation $y' = y^2/3$, $y(0) = 1$. Let us try to approximate $y(2)$ using Euler's method. In Figures 1.9 and 1.10 we have graphically approximated $y(2)$ with step size 1. With step size 1, we have $y(2) \approx 1.926$. The real answer is 3. We are approximately 1.074 off. Let us halve the step size. Computing y_4 with $h = 0.5$, we find that $y(2) \approx 2.209$, so an error of about 0.791. Table 1.1 on the next page gives the values computed for various parameters.

Exercise 1.6.1: Solve this equation exactly and show that $y(2) = 3$.

The difference between the actual solution and the approximate solution is called the error. We usually talk about just the size of the error and we do not care much about its sign. The point is, we usually do not know the real solution, so we only have a vague understanding of the error. If we knew the error exactly ... what is the point of doing the approximation?

h	Approximate $y(2)$	Error	$\frac{\text{Error}}{\text{Previous error}}$
1	1.92593	1.07407	
0.5	2.20861	0.79139	0.73681
0.25	2.47250	0.52751	0.66656
0.125	2.68034	0.31966	0.60599
0.0625	2.82040	0.17960	0.56184
0.03125	2.90412	0.09588	0.53385
0.015625	2.95035	0.04965	0.51779
0.0078125	2.97472	0.02528	0.50913

Table 1.1: Euler's method approximation of $y(2)$ where of $y' = y^2/3$, $y(0) = 1$.

Notice that except for the first few times, every time we halved the interval the error approximately halved. This halving of the error is a general feature of Euler's method as it is a *first order method*. There exists an improved Euler method, see the exercises, which is a second order method. A second order method reduces the error to approximately one quarter every time we halve the interval. The meaning of “second” order is the squaring in $1/4 = 1/2 \times 1/2 = (1/2)^2$.

To get the error to be within 0.1 of the answer we had to already do 64 steps. To get it to within 0.01 we would have to halve another three or four times, meaning doing 512 to 1024 steps. That is quite a bit to do by hand. The improved Euler method from the exercises should quarter the error every time we halve the interval, so we would have to approximately do half as many “halvings” to get the same error. This reduction can be a big deal. With 10 halvings (starting at $h = 1$) we have 1024 steps, whereas with 5 halvings we only have to do 32 steps, assuming that the error was comparable to start with. A computer may not care about this difference for a problem this simple, but suppose each step would take a second to compute (the function may be substantially more difficult to compute than $y^2/3$). Then the difference is 32 seconds versus about 17 minutes. We are not being altogether fair, a second order method would probably double the time to do each step. Even so, it is 1 minute versus 17 minutes. Next, suppose that we have to repeat such a calculation for different parameters a thousand times. You get the idea.

Note that in practice we do not know how large the error is! How do we know what is the right step size? Well, essentially we keep halving the interval, and if we are lucky, we can estimate the error from a few of these calculations and the assumption that the error goes down by a factor of one half each time (if we are using standard Euler).

Exercise 1.6.2: In the table above, suppose you do not know the error. Take the approximate values of the function in the last two lines, assume that the error goes down by a factor of 2. Can you estimate the error in the last time from this? Does it (approximately) agree with the table? Now do it for the first two rows. Does this agree with the table?

Let us talk a little bit more about the example $y' = \frac{y^2}{3}$, $y(0) = 1$. Suppose that instead

of the value $y(2)$ we wish to find $y(3)$. The results of this effort are listed in Table 1.2 for successive halvings of h . What is going on here? Well, you should solve the equation exactly and you will notice that the solution does not exist at $x = 3$. In fact, the solution goes to infinity when you approach $x = 3$.

h	Approximate $y(3)$
1	3.16232
0.5	4.54329
0.25	6.86079
0.125	10.80321
0.0625	17.59893
0.03125	29.46004
0.015625	50.40121
0.0078125	87.75769

Table 1.2: Attempts to use Euler's to approximate $y(3)$ where of $y' = y^2/3$, $y(0) = 1$.

Another case where things go bad is if the solution oscillates wildly near some point. The solution may exist at all points, but even a much better numerical method than Euler would need an insanely small step size to approximate the solution with reasonable precision. And computers might not be able to easily handle such a small step size.

In real applications we would not use a simple method such as Euler's. The simplest method that would probably be used in a real application is the standard Runge–Kutta method (see exercises). That is a fourth order method, meaning that if we halve the interval, the error generally goes down by a factor of 16 (it is fourth order as $1/16 = 1/2 \times 1/2 \times 1/2 \times 1/2$).

Choosing the right method to use and the right step size can be very tricky. There are several competing factors to consider.

- Computational time: Each step takes computer time. Even if the function f is simple to compute, we do it many times over. Large step size means faster computation, but perhaps not the right precision.
- Roundoff errors: Computers only compute with a certain number of significant digits. Errors introduced by rounding numbers off during our computations become noticeable when the step size becomes too small relative to the quantities we are working with. So reducing step size may in fact make errors worse. There is a certain optimum step size such that the precision increases as we approach it, but then starts getting worse as we make our step size smaller still. Trouble is: this optimum may be hard to find.
- Stability: Certain equations may be numerically unstable. What may happen is that the numbers never seem to stabilize no matter how many times we halve the interval. We may need a ridiculously small interval size, which may not be practical due to roundoff

errors or computational time considerations. Such problems are sometimes called *stiff*. In the worst case, the numerical computations might be giving us bogus numbers that look like a correct answer. Just because the numbers seem to have stabilized after successive halving, does not mean that we must have the right answer.

We have seen just the beginnings of the challenges that appear in real applications. Numerical approximation of solutions to differential equations is an active research area for engineers and mathematicians. For example, the general purpose method used for the ODE solver in Matlab and Octave (as of this writing) is a method that appeared in the literature only in the 1980s.

The method used in Matlab and Octave is a fair bit different from the methods discussed previously. We don't need to go too much in detail about it, but some information will be helpful. The main difference that will be visible when running these methods is that they are *adaptive* method. This means that they adjust the step-size based on what the differential equation looks like at a given point. Euler's method, along with the improved Euler and Runge-Kutta methods, is a fixed step-size method, where the steps are always the same no matter what. Adaptive methods are harder to write and optimize, but can solve many problems faster because the adaptive nature of the method allows them to get similar accuracy to fixed step methods, but at many fewer steps. In the example below, the initial value problem

$$\frac{dy}{dx} = y \quad y(0) = 1$$

is solved with an Euler's method and Matlab's built-in `ode45` method. Both of the solutions are plotted along with the actual solution $y = e^x$

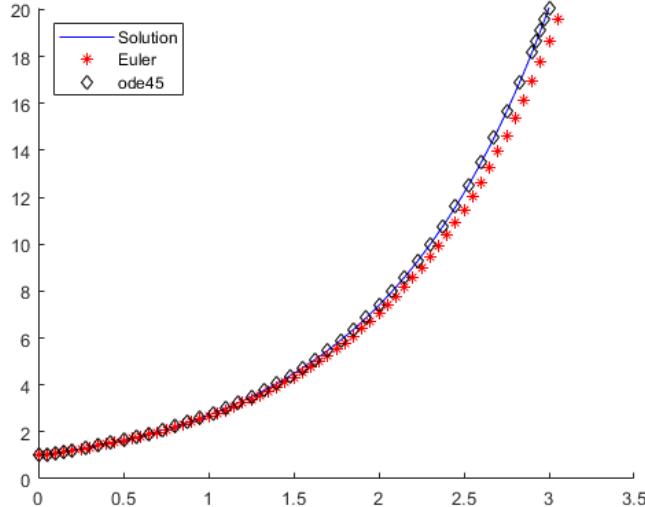


Figure 1.11: Comparison of the solution from Euler's Method and `ode45` to the actual solution of $\frac{dy}{dx} = y$.

The Euler's method takes 60 steps in this computation, but is still not as accurate as the `ode45` method, which only takes 45 steps. In addition, the black diamonds, representing the different values computed by `ode45` are not evenly spaced, illustrating the adaptive nature of this solver, while the red stars are all evenly-spaced, as is expected from Euler's method.

1.6.1 Exercises

Exercise 1.6.3: Consider $\frac{dx}{dt} = (2t - x)^2$, $x(0) = 2$. Use Euler's method with step size $h = 0.5$ to approximate $x(1)$.

Exercise 1.6.4: Consider the differential equation $\frac{dy}{dt} = t^2 - 3y + 1$ with $y(1) = 4$. Approximate the solution at $t = 3$ using Euler's method with a step size of $h = 1$ and $h = 0.5$. Compare these values with the actual solution at $t = 3$.

Exercise 1.6.5: Consider the differential equation $\frac{dy}{dt} = 2ty + y^2$ with $y(0) = 1$. Approximate the solution at $t = 2$ using Euler's method with a step size of $h = 1$ and $h = 0.5$.

Exercise 1.6.6: Consider $\frac{dx}{dt} = t - x$, $x(0) = 1$.

- a) Use Euler's method with step sizes $h = 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$ to approximate $x(1)$.
- b) Solve the equation exactly.
- c) Describe what happens to the errors for each h you used. That is, find the factor by which the error changed each time you halved the interval.

Exercise 1.6.7:* Let $x' = \sin(xt)$, and $x(0) = 1$. Approximate $x(1)$ using Euler's method with step sizes 1, 0.5, 0.25. Use a calculator and compute up to 4 decimal digits.

Exercise 1.6.8: Approximate the value of e by looking at the initial value problem $y' = y$ with $y(0) = 1$ and approximating $y(1)$ using Euler's method with a step size of 0.2.

Exercise 1.6.9:* Let $x' = 2t$, and $x(0) = 0$.

- a) Approximate $x(4)$ using Euler's method with step sizes 4, 2, and 1.
- b) Solve exactly, and compute the errors.
- c) Compute the factor by which the errors changed.

Exercise 1.6.10:* Let $x' = xe^{xt+1}$, and $x(0) = 0$.

- a) Approximate $x(4)$ using Euler's method with step sizes 4, 2, and 1.
- b) Guess an exact solution based on part a) and compute the errors.

Exercise 1.6.11: Example of numerical instability: Take $y' = -5y$, $y(0) = 1$. We know that the solution should decay to zero as x grows. Using Euler's method, start with $h = 1$ and compute y_1, y_2, y_3, y_4 to try to approximate $y(4)$. What happened? Now halve the interval. Keep halving the interval and approximating $y(4)$ until the numbers you are getting start to stabilize (that is, until they start going towards zero). Note: You might want to use a calculator.

There is a simple way to improve Euler's method to make it a second order method by doing just one extra step. Consider $\frac{dy}{dx} = f(x, y)$, $y(x_0) = y_0$, and a step size h . What we do is to pretend we compute the next step as in Euler, that is, we start with (x_i, y_i) , we compute a slope $k_1 = f(x_i, y_i)$, and then look at the point $(x_i + h, y_i + k_1 h)$. Instead of letting our new point be $(x_i + h, y_i + k_1 h)$, we compute the slope at that point, call it k_2 , and then take the average of k_1 and k_2 , hoping that the average is going to be closer to the actual slope on the interval from x_i to $x_i + h$. And we are correct, if we halve the step, the error should go down by a factor of $2^2 = 4$. To summarize, the setup is the same as for regular Euler, except the computation of y_{i+1} and x_{i+1} .

$$\begin{aligned} k_1 &= f(x_i, y_i), & x_{i+1} &= x_i + h, \\ k_2 &= f(x_i + h, y_i + k_1 h), & y_{i+1} &= y_i + \frac{k_1 + k_2}{2} h. \end{aligned}$$

Exercise 1.6.12:* Consider $\frac{dy}{dx} = x + y$, $y(0) = 1$.

- a) Use the improved Euler's method (see above) with step sizes $h = 1/4$ and $h = 1/8$ to approximate $y(1)$.
- b) Use Euler's method with $h = 1/4$ and $h = 1/8$.
- c) Solve exactly, find the exact value of $y(1)$.
- d) Compute the errors, and the factors by which the errors changed.

The simplest method used in practice is the *Runge–Kutta method*. Consider $\frac{dy}{dx} = f(x, y)$, $y(x_0) = y_0$, and a step size h . Everything is the same as in Euler's method, except the computation of y_{i+1} and x_{i+1} .

$$\begin{aligned} k_1 &= f(x_i, y_i), & x_{i+1} &= x_i + h, \\ k_2 &= f(x_i + h/2, y_i + k_1(h/2)), & y_{i+1} &= y_i + \frac{k_1 + 2k_2 + 2k_3 + k_4}{6} h, \\ k_3 &= f(x_i + h/2, y_i + k_2(h/2)), \\ k_4 &= f(x_i + h, y_i + k_3 h). \end{aligned}$$

Exercise 1.6.13: Consider $\frac{dy}{dx} = yx^2$, $y(0) = 1$.

- a) Use Runge–Kutta (see above) with step sizes $h = 1$ and $h = 1/2$ to approximate $y(1)$.
- b) Use Euler's method with $h = 1$ and $h = 1/2$.
- c) Solve exactly, find the exact value of $y(1)$, and compare.

1.7 Autonomous equations

Attribution: [JL], §1.6.

Learning Objectives

After this section, you will be able to:

- Identify autonomous first order differential equations,
- Find critical points or equilibrium solutions for autonomous equations,
- Sketch a phase line for these equations, and
- Draw and analyze bifurcation diagrams for autonomous equations with parameter.

Definition 1.7.1

An equation of the form

$$\frac{dx}{dt} = f(x),$$

where the derivative of solutions depends only on x (the dependent variable) is called an *autonomous equation*.

If we think of t as time, the naming comes from the fact that the equation is independent of time.

We return to the cooling coffee problem (Example 1.3.3). Newton's law of cooling says

$$\frac{dx}{dt} = k(A - x),$$

where x is the temperature, t is time, k is some positive constant, and A is the ambient temperature. See Figure 1.12 on the next page for an example with $k = 0.3$ and $A = 5$.

Note the solution $x = A$ (in the figure $x = 5$). We call these constant solutions the *equilibrium solutions*. The points on the x -axis where $f(x) = 0$ are called *critical points*. The point $x = A$ is a critical point. In fact, each critical point corresponds to an equilibrium solution.

Now, we want to determine what happens for other values of x that are not A . Based on the existence and uniqueness theorem in § 1.5 for first order differential equations, the fact that $k(A - x)$ and its partial derivative $-k$ are continuous everywhere gives that solution curves can not cross. This means that since we know $x = A$ is a solution, if a solution starts below $x = A$, it must always stay there, and solutions that start above $x = A$ will also stay there. For more information about what the solutions do, we'll need to look back at the equation and some sample solution curves.

Note also, by looking at the graph, that the solution $x = A$ is “stable” in that small perturbations in x do not lead to substantially different solutions as t grows. If we change the initial condition a little bit, then as $t \rightarrow \infty$ we get $x(t) \rightarrow A$. We call such a critical point *asymptotically stable*. In this simple example it turns out that all solutions in fact go to A as $t \rightarrow \infty$. If there is a critical point where all nearby solutions move away from the

critical point, we say it is *unstable*. If some nearby solutions go towards the critical point, and some others move away, then we say it is *semistable*. The final option is that solutions nearby neither move towards nor away from the critical point, and these critical points are called *stable*.

The last of these options may seem strange at first, and that is because stable critical points are not possible for autonomous equations with one unknown function. If a solution does not move towards or away from a critical point, that means it doesn't move anywhere, and so is a critical point on its own. However, when we get to autonomous systems in § 4.5 and § 5.2, we will see that in two dimensions, this is possible (think of a circle that does not spiral into or away from the center point).

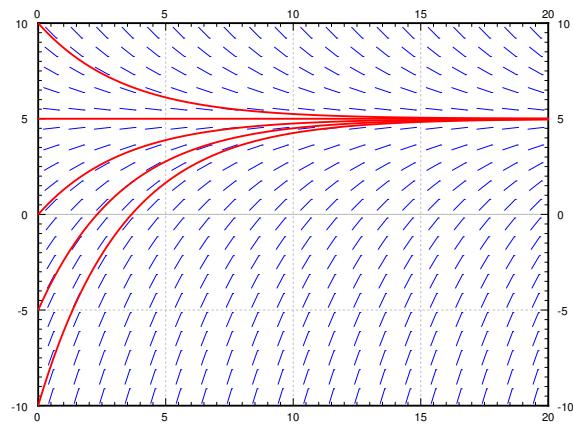


Figure 1.12: The slope field and some solutions of $x' = 0.3(5 - x)$.

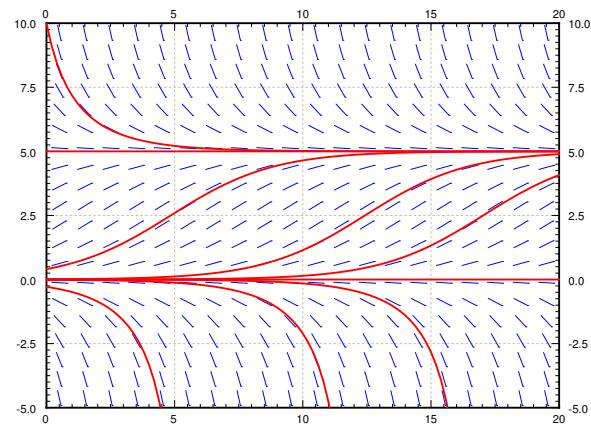


Figure 1.13: The slope field and some solutions of $x' = 0.1x(5 - x)$.

Consider now the *logistic equation*

$$\frac{dx}{dt} = kx(M - x),$$

for some positive k and M . This equation is commonly used to model population if we know the limiting population M , that is the maximum sustainable population. The logistic equation leads to less catastrophic predictions on world population than $x' = kx$. In the real world there is no such thing as negative population, but we will still consider negative x for the purposes of the math.

See Figure 1.13 for an example, $x' = 0.1x(5 - x)$. There are two critical points, $x = 0$ and $x = 5$. The critical point at $x = 5$ is asymptotically stable, while the critical point at $x = 0$ is unstable.

It is not necessary to find the exact solutions to talk about the long term behavior of the solutions. From the slope field above of $x' = 0.1x(5 - x)$, we see that

$$\lim_{t \rightarrow \infty} x(t) = \begin{cases} 5 & \text{if } x(0) > 0, \\ 0 & \text{if } x(0) = 0, \\ \text{DNE or } -\infty & \text{if } x(0) < 0. \end{cases}$$

Here DNE means “does not exist.” From just looking at the slope field we cannot quite decide what happens if $x(0) < 0$. It could be that the solution does not exist for t all the way to ∞ . Think of the equation $x' = x^2$; we have seen that solutions only exist for some finite period of time. Same can happen here. In our example equation above it turns out that the solution does not exist for all time, but to see that we would have to solve the equation. In any case, the solution does go to $-\infty$, but it may get there rather quickly.

If we are interested only in the long term behavior of the solution, we would be doing unnecessary work if we solved the equation exactly. We could draw the slope field, but it is easier to just look at the *phase diagram* or *phase line*, which is a simple way to visualize the behavior of autonomous equations. The phase line for this equation is visible in [Figure 1.14](#). In this case there is one dependent variable x . We draw the x -axis, we mark all the critical points, and then we draw arrows in between. Since x is the dependent variable we draw the axis vertically, as it appears in the slope field diagrams above. If $f(x) > 0$, we draw an up arrow. If $f(x) < 0$, we draw a down arrow. To figure this out, we could just plug in some x between the critical points, $f(x)$ will have the same sign at all x between two critical points as long $f(x)$ is continuous. For example, $f(6) = -0.6 < 0$, so $f(x) < 0$ for $x > 5$, and the arrow above $x = 5$ is a down arrow. Next, $f(1) = 0.4 > 0$, so $f(x) > 0$ whenever $0 < x < 5$, and the arrow points up. Finally, $f(-1) = -0.6 < 0$ so $f(x) < 0$ when $x < 0$, and the arrow points down.

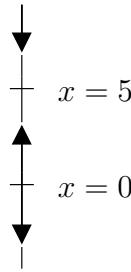


Figure 1.14: Phase line for the differential equation $x' = 0.1x(5 - x)$.

Armed with the phase diagram, it is easy to sketch the solutions approximately: As time t moves from left to right, the graph of a solution goes up if the arrow is up, and it goes down if the arrow is down.

Exercise 1.7.1: Try sketching a few solutions simply from looking at the phase diagram. Check with the preceding graphs if you are getting the type of curves.

Once we draw the phase diagram, we classify critical points as asymptotically stable, semistable, or unstable based on whether the “arrows” point into or away from the critical point on each side. Two arrows in means that the critical point is asymptotically stable, two arrows away means unstable, and one in one out means semistable.

Example 1.7.1: Consider the autonomous differential equation

$$\frac{dx}{dt} = x(x - 2)^2(x + 3)(x - 4) \quad (1.5)$$

Find all equilibrium solutions for this equation, and determine their stability. Draw a phase line and use this information to sketch some approximate solution curves.

Solution: This equation is already in factored form. This makes it simple to determine the equilibrium solutions as $x = 0$, $x = 2$, $x = -3$ and $x = 4$. In order to determine the stability of each critical point and draw the phase line, we need to plug in values surrounding these points to $f(x) = x(x - 2)^2(x + 3)(x - 4)$. We can see that

$$\begin{aligned}f(-4) &= (-4)(-6)^2(-1)(-8) < 0, \\f(-1) &= (-1)(-3)^2(2)(-5) > 0, \\f(1) &= (1)(-1)^2(4)(-3) < 0, \\f(3) &= (3)(1)^2(6)(-1) < 0, \\f(5) &= (5)(3)^2(8)(1) > 0.\end{aligned}$$

This lets us draw the phase line and determine the stability of each critical point. Thus, we see that $x = -3$ is an unstable critical point, $x = 0$ is asymptotically stable, $x = 2$ is semistable, and $x = 4$ is unstable. A set of sample solution curves also validates these conclusions.

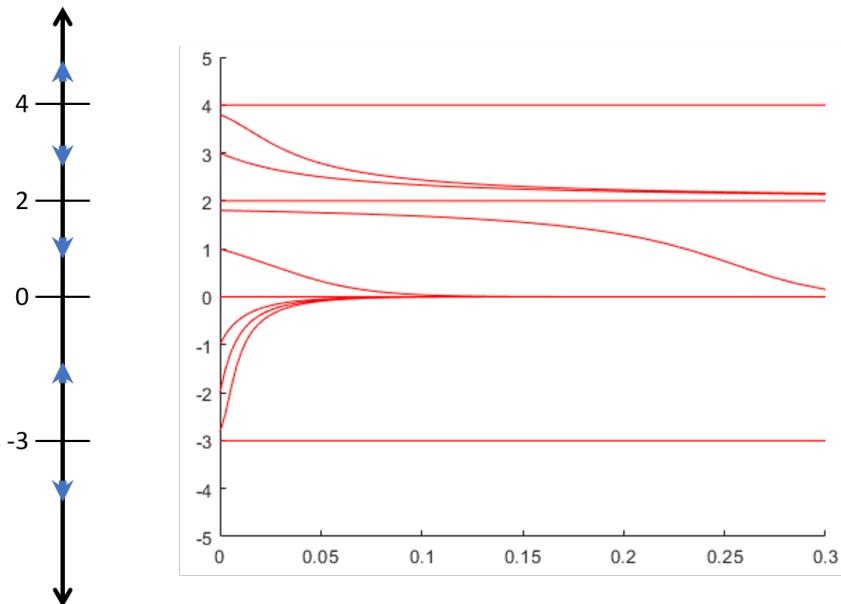


Figure 1.15: Phase line for the differential equation $\frac{dx}{dt} = x(x - 2)^2(x + 3)(x - 4)$ and a plot of some solutions to this equation.

1.7.1 Concavity of Solutions

We can tell from the phase line for an autonomous equation when the solution will be increasing or decreasing. Is there any more we can learn about the shape of these graphs?

There is, and it comes from looking for the concavity, which is determined by the second derivative.

We can compute the second derivative

$$\frac{d^2x}{dt^2} = \frac{d}{dx} \left[\frac{dx}{dt} \right]$$

of our solution by noticing that $\frac{dx}{dt} = f(x)$. This function can be differentiated by the chain rule

$$\frac{d}{dt} f(x) = f'(x) \frac{dx}{dt} = f'(x)f(x).$$

So, the solution is concave up if $f'(x)f(x)$ is positive, and concave down if that is negative. Phrased another way, the solution is concave up if f and f' have the same sign, and it is concave down if f and f' have opposite signs.

Let's see what this looks like in action. Take the logistic equation $x' = 0.1x(5 - x)$, whose solutions are plotted in [Figure 1.13](#). [Figure 1.16](#) shows the graph of $f(x)$ as a function of x for this scenario. When do f and f' have the same sign? Well, this happens when f is both positive and increasing, or negative and decreasing. This happens between 0 and the vertex, as well as for $x > 5$. The vertex here is at $x = 2.5$, and so we conclude that the solution should be concave up when x is on the intervals $(0, 2.5)$ and $(5, \infty)$, and be concave down otherwise. Looking back at [Figure 1.13](#), this is exactly what we observe. All of the solutions between 0 and 5 seem to "flip over" to be concave down when x crosses 2.5.

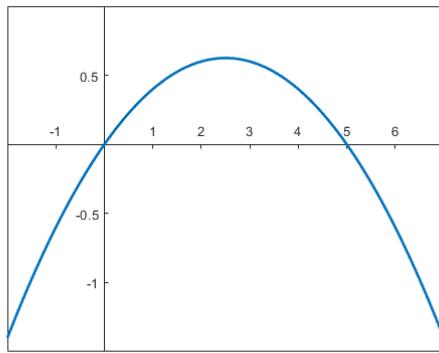


Figure 1.16: Plot of x vs. $f(x)$ for the differential equation $\frac{dx}{dt} = 0.1x(5 - x)$.

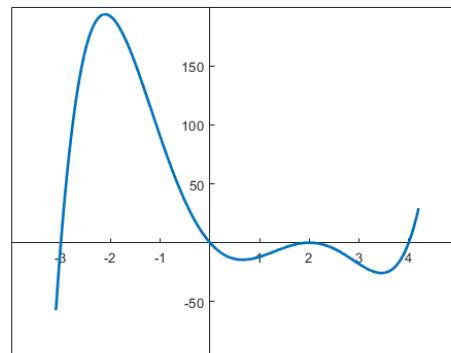


Figure 1.17: Plot of x vs. $f(x)$ for the differential equation $\frac{dx}{dt} = x(x-2)^2(x+3)(x-4)$.

The same can be seen for solutions to [\(1.5\)](#), even though we can't compute the extreme values explicitly. [Figure 1.17](#) shows the graph of $f(x)$ vs. x for this situation. Between each pair of equilibrium solutions there is a critical point of f (in the Calculus 1 sense) where the derivative is zero, and at this point, the derivative changes sign, and since the function value does not change sign, the concavity of the solution to the differential equation flips at this point. Comparing this graph and these points where concavity shifts with the solutions drawn in [Figure 1.15](#) again validates these results.

1.7.2 Bifurcation Diagrams

An extension of the topic of autonomous equation is *autonomous equations with parameter*. The idea is that we have a differential equation that has no explicit dependence on time, but does have a dependence on an outside parameter, which is a constant set by the physical situation. In terms of physical problems, this parameter will tend to be something that we can adjust to change how the differential equation behaves. For example, in a logistic differential equation

$$\frac{dx}{dt} = ax(K - x)$$

both the a or K could be adjustable parameters. For a given value of the parameter, the differential equation behaves like a standard autonomous differential equation, but we can get different properties of this equation for different values of the parameter.

Definition 1.7.2

An *autonomous equation with parameter α* is a differential equation of the form

$$\frac{dx}{dt} = f_\alpha(x)$$

where, for every value of α , $f_\alpha(x)$ is a function of the single variable x .

Later, we will want to view $f_\alpha(x)$ as a two-variable function of x and α , but for now, we want to think about it as a function of just x for a fixed value of α . We want to be able to analyze what happens to this equation for different values of α . Since it is an autonomous equation, we can do this using phase lines. This will be easiest to see through an example.

Example 1.7.2: Consider the differential equation

$$\frac{dx}{dt} = x(x^2 - \alpha),$$

which fits the description of an autonomous equation with parameter α . We want to see what happens for different values of α .

Solution: We can draw a phase line for $\alpha = -4$, $\alpha = 0$ and $\alpha = 1$. It is clear that something happens with this equation between $\alpha = -4$ and $\alpha = 1$. We go from having only one equilibrium solution at $\alpha = -4$ to having three equilibrium solutions at $\alpha = 1$. In addition, the solution at $y = 0$ is unstable for $\alpha = -4$, while it is asymptotically stable for $\alpha = 1$. If we want to figure out when this change happens, we'll need a better way to analyze this problem. \square

How can we better approach this problem? The idea is to think about when the solution to the differential equation will be increasing or decreasing as a function of the two variables α and x . Since a phase line is a plot of when the solution is increasing or decreasing for a given value of α , we essentially want to plot all of these phase lines on a two-dimensional graph. This graph is called a *bifurcation diagram*. Figure 1.19 on the following page shows a bifurcation diagram for the example $\frac{dx}{dt} = x(x^2 - \alpha)$.

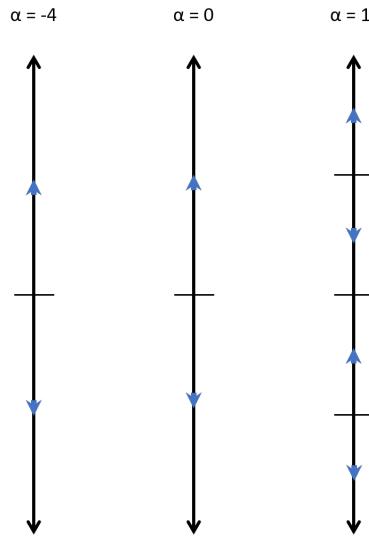


Figure 1.18: Phase lines for the differential equation $\frac{dx}{dt} = x(x^2 - \alpha)$ for $\alpha = -4, 0, 1$.

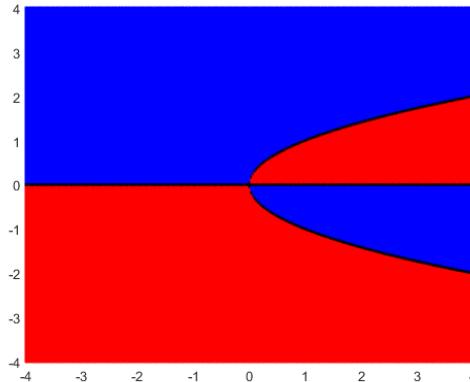


Figure 1.19: Bifurcation Diagram for the differential equation $\frac{dx}{dt} = x(x^2 - \alpha)$. In this figure, a blue region means the solution will be increasing and red indicates decreasing.

Within this picture, we can see all of our phase lines from before, because at any value of α , taking the vertical slice of this graph at that value, we get the phase line. If we want to consider $\alpha = -4$, then we can look above the horizontal coordinate -4 , and that will give us the phase line for $\alpha = -4$. The same goes for any other value of α we want to consider. For instance, we can also see that for any $\alpha \leq 0$, there will be one equilibrium solution, and for $\alpha > 0$ there are three equilibrium solutions, indicated by the three black curves above each of those α values.

From this, we can see that the point at which the behavior changes is $\alpha = 0$. Thus, for this problem $\alpha = 0$ is called the bifurcation point. This is defined to be the value of the

parameter for which the overall behavior of the equation changes. This can be a change in the number of equilibrium solutions, the stability of these equilibrium solutions, or both. For this particular example, we have both of these. We go from 1 equilibrium solution to 3, and the solution at $y = 0$ changes in stability. This type of bifurcation is called a “pitchfork bifurcation” based on the shape of the equilibrium solutions near the bifurcation point.

Another example of a bifurcation of a different form can be seen in the example of the logistic equation with harvesting. Suppose an alien race really likes to eat humans. They keep a planet with humans on it and harvest the humans at a rate of h million humans per year. Suppose x is the number of humans in millions on the planet and t is time in years. Let M be the limiting population when no harvesting is done. The number $k > 0$ is a constant depending on how fast humans multiply. Our equation becomes

$$\frac{dx}{dt} = kx(M - x) - h.$$

In this setup, M and k are fixed values, and the parameter that is being adjusted for this equation is h . We expand the right-hand side and set it to zero.

$$kx(M - x) - h = -kx^2 + kMx - h = 0.$$

Solving for the critical points using the quadratic formula, let us call them A and B , we get

$$A = \frac{kM + \sqrt{(kM)^2 - 4hk}}{2k}, \quad B = \frac{kM - \sqrt{(kM)^2 - 4hk}}{2k}.$$

Exercise 1.7.2: Sketch a phase diagram for different possibilities. Note that these possibilities are $A > B$, or $A = B$, or A and B both complex (i.e. no real solutions). Hint: Fix some simple k and M and then vary h .

Example 1.7.3: For example, let $M = 8$ and $k = 0.1$. What happens for different values of h in this situation?

Solution: When $h = 1$, then A and B are distinct and positive. The slope field we get is in [Figure 1.20](#) on the next page. As long as the population starts above B , which is approximately 1.55 million, then the population will not die out. It will in fact tend towards $A \approx 6.45$ million. If ever some catastrophe happens and the population drops below B , humans will die out, and the fast food restaurant serving them will go out of business.

When $h = 1.6$, then $A = B = 4$. There is only one critical point and it is semistable. When the population starts above 4 million it will tend towards 4 million. If it ever drops below 4 million, humans will die out on the planet. This scenario is not one that we (as the human fast food proprietor) want to be in. A small perturbation of the equilibrium state and we are out of business. There is no room for error. See [Figure 1.21](#) on the following page.

Finally if we are harvesting at 2 million humans per year, there are no critical points. The population will always plummet towards zero, no matter how well stocked the planet starts. See [Figure 1.22](#) on the next page.

All of these can also be seen from the bifurcation diagram, which is drawn in [Figure 1.23](#) on page 69. The values A and B discussed above represent the upper and lower branches of

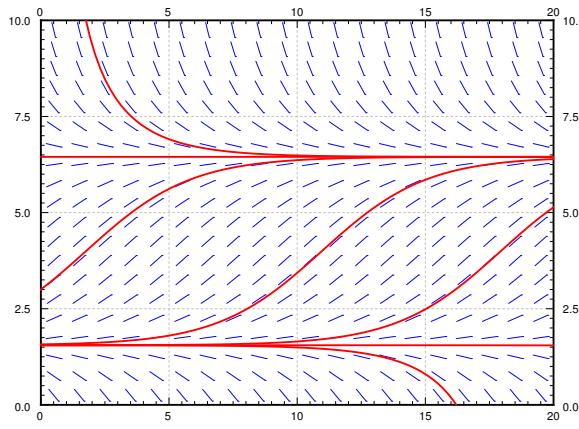


Figure 1.20: The slope field and some solutions of $x' = 0.1x(8 - x) - 1$.

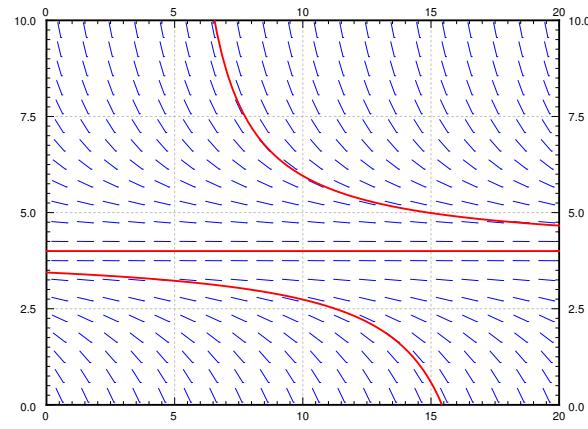


Figure 1.21: The slope field and some solutions of $x' = 0.1x(8 - x) - 1.6$.

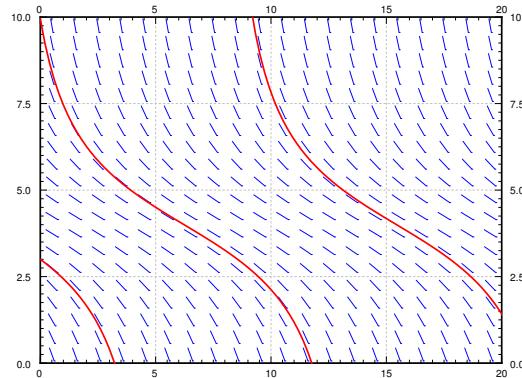


Figure 1.22: The slope field and some solutions of $x' = 0.1x(8 - x) - 2$.

the parabola in the figure. For any $h > 1.6$, there are no equilibrium solutions and the phase line is entirely decreasing, meaning the solution will converge to zero no matter what. For $h < 1.6$, there are two equilibrium solutions, with the top one asymptotically stable and the bottom one unstable. At $h = 1.6$ is where the bifurcation point occurs for this example. This is an example of a “saddle-node” bifurcation, as the two equilibrium solutions collide with each other at the bifurcation point and disappear.

Another way to visualize this situation is by plotting the function $f_\alpha(x)$ for the different values of α . The places where this function is zero give the equilibrium solutions, and we can determine *bifurcation values* by looking for where the zeros of this function change behavior. For this particular example, the graphs of $f_\alpha(x)$ are drawn in Figure 1.24 on the next page.

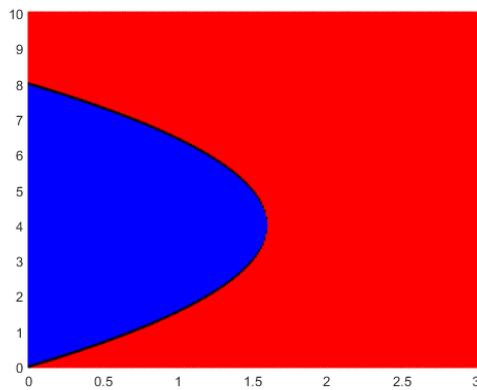


Figure 1.23: Bifurcation diagram for the differential equation $x' = 0.1x(8 - x) - h$.

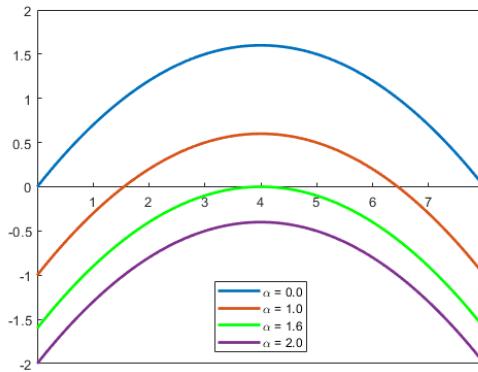


Figure 1.24: Graph of $f_\alpha(x) = 0.1x(8 - x) - \alpha$ for $\alpha = 0, 1.0, 1.6, 2.0$.

1.7.3 Exercises

Exercise 1.7.3: Consider $x' = x^2$.

- Draw the phase diagram, find the critical points, and mark them asymptotically stable, semistable, or unstable.
- Sketch typical solutions of the equation.
- Find $\lim_{t \rightarrow \infty} x(t)$ for the solution with the initial condition $x(0) = -1$.

Exercise 1.7.4: Consider $x' = \sin x$.

- Draw the phase diagram for $-4\pi \leq x \leq 4\pi$. On this interval mark the critical points asymptotically stable, semistable, or unstable.

- b) Sketch typical solutions of the equation.
 c) Find $\lim_{t \rightarrow \infty} x(t)$ for the solution with the initial condition $x(0) = 1$.

Exercise 1.7.5:* Let $x' = (x - 1)(x - 2)x^2$.

- a) Sketch the phase diagram and find critical points.
 b) Classify the critical points.
 c) If $x(0) = 0.5$, then find $\lim_{t \rightarrow \infty} x(t)$.

Exercise 1.7.6: Let $y' = (y - 2)(y^2 + 1)(y + 3)$. Sketch a phase diagram for this differential equation. Find and classify all critical points. If $y(0) = 0$, what will happen to the solution as $t \rightarrow \infty$?

Exercise 1.7.7: Find and classify all equilibrium solutions for the differential equation $x' = (x - 2)^2(x + 1)(x + 3)^3(x + 2)$.

Exercise 1.7.8:* Let $x' = e^{-x}$.

- a) Find and classify all critical points. b) Find $\lim_{t \rightarrow \infty} x(t)$ given any initial condition.

Exercise 1.7.9: Suppose $f(x)$ is positive for $0 < x < 1$, it is zero when $x = 0$ and $x = 1$, and it is negative for all other x .

- a) Draw the phase diagram for $x' = f(x)$, find the critical points, and mark them asymptotically stable, semistable, or unstable.
 b) Sketch typical solutions of the equation.
 c) Find $\lim_{t \rightarrow \infty} x(t)$ for the solution with the initial condition $x(0) = 0.5$.

Exercise 1.7.10:* Suppose $\frac{dx}{dt} = (x - \alpha)(x - \beta)$ for two numbers $\alpha < \beta$.

- a) Find the critical points, and classify them.

For b), c), d), find $\lim_{t \rightarrow \infty} x(t)$ based on the phase diagram.

- b) $x(0) < \alpha$, c) $\alpha < x(0) < \beta$, d) $\beta < x(0)$.

Exercise 1.7.11: Start with the logistic equation $\frac{dx}{dt} = kx(M - x)$. Suppose we modify our harvesting. That is we will only harvest an amount proportional to current population. In other words, we harvest hx per unit of time for some $h > 0$ (Similar to earlier example with h replaced with hx).

- a) Construct the differential equation.
 b) Show that if $kM > h$, then the equation is still logistic.
 c) What happens when $kM < h$?

Exercise 1.7.12:* Assume that a population of fish in a lake satisfies $\frac{dx}{dt} = kx(M - x)$. Now suppose that fish are continually added at A fish per unit of time.

- a) Find the differential equation for x . b) What is the new limiting population?

Exercise 1.7.13: A disease is spreading through the country. Let x be the number of people infected. Let the constant S be the number of people susceptible to infection. The infection rate $\frac{dx}{dt}$ is proportional to the product of already infected people, x , and the number of susceptible but uninfected people, $S - x$.

- a) Write down the differential equation.
 b) Supposing $x(0) > 0$, that is, some people are infected at time $t = 0$, what is $\lim_{t \rightarrow \infty} x(t)$.
 c) Does the solution to part b) agree with your intuition? Why or why not?

Exercise 1.7.14: Consider the differential equation with parameter α given by $y' = y(y - \alpha + 1)$.

- a) Sketch a phase diagram for this differential equation with $\alpha = -3$, $\alpha = 1$, and $\alpha = 3$.
 b) Draw a bifurcation diagram for this differential equation with parameter.

What is the bifurcation point for this equation? What changes when α passes over the bifurcation point?

Exercise 1.7.15: Consider the differential equation with parameter α given by $y' = y^2(y^2 - \alpha)$.

- a) Sketch a phase diagram for this differential equation with $\alpha = -3$, $\alpha = 0$, and $\alpha = 3$.
 b) Draw a bifurcation diagram for this differential equation with parameter.

What is the bifurcation point for this equation? What changes when α passes over the bifurcation point?

1.8 Exact equations

Attribution: [JL], §1.8.

Learning Objectives

After this section, you will be able to:

- Determine if a first order differential equation is exact,
- Find the general solution to an exact equation,
- Solve initial value problems for exact equations, and
- Use integrating factors to make some non-exact equations exact in order to solve them.

Another type of equation that comes up quite often in physics and engineering is an *exact equation*. Suppose $F(x, y)$ is a function of two variables, which we call the *potential function*. The naming should suggest potential energy, or electric potential. Exact equations and potential functions appear when there is a conservation law at play, such as conservation of energy. Let us make up a simple example. Let

$$F(x, y) = x^2 + y^2.$$

We are interested in the lines of constant energy, that is lines where the energy is conserved; we want curves where $F(x, y) = C$, for some constant C . In our example, the curves $x^2 + y^2 = C$ are circles. See Figure 1.25.

We take the *total derivative* of F :

$$dF = \frac{\partial F}{\partial x} dx + \frac{\partial F}{\partial y} dy.$$

For convenience, we will make use of the notation of $F_x = \frac{\partial F}{\partial x}$ and $F_y = \frac{\partial F}{\partial y}$. In our example,

$$dF = 2x dx + 2y dy.$$

We apply the total derivative to $F(x, y) = C$, to find the differential equation $dF = 0$. The differential equation we obtain in such a way has the form

$$M dx + N dy = 0, \quad \text{or} \quad M + N \frac{dy}{dx} = 0.$$

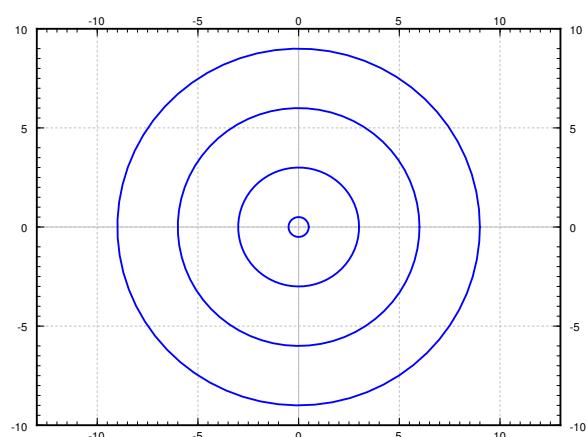


Figure 1.25: Solutions to $F(x, y) = x^2 + y^2 = C$ for various C .

Definition 1.8.1

An equation of the form

$$M(x, y) + N(x, y) \frac{dy}{dx} = 0$$

is called *exact* if it was obtained as $dF = 0$ for some potential function F .

In our simple example, we obtain the equation

$$2x \, dx + 2y \, dy = 0, \quad \text{or} \quad 2x + 2y \frac{dy}{dx} = 0.$$

Since we obtained this equation by differentiating $x^2 + y^2 = C$, the equation is exact. We often wish to solve for y in terms of x . In our example,

$$y = \pm \sqrt{C^2 - x^2}.$$

An interpretation of the setup is that at each point $\vec{v} = (M, N)$ is a vector in the plane, that is, a direction and a magnitude. As M and N are functions of (x, y) , we have a *vector field*. The particular vector field \vec{v} that comes from an exact equation is a so-called *conservative vector field*, that is, a vector field that comes with a potential function $F(x, y)$, such that

$$\vec{v} = \left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y} \right).$$

As we will see shortly, the process of solving an exact equation is basically identical to the process of finding a potential function from a conservative vector field. Let γ be a path in the plane starting at (x_1, y_1) and ending at (x_2, y_2) . If we think of \vec{v} as force, then the work required to move along γ is

$$\int_{\gamma} \vec{v}(\vec{r}) \cdot d\vec{r} = \int_{\gamma} M \, dx + N \, dy = F(x_2, y_2) - F(x_1, y_1).$$

That is, the work done only depends on endpoints, that is where we start and where we end. For example, suppose F is gravitational potential. The derivative of F given by \vec{v} is the gravitational force. What we are saying is that the work required to move a heavy box from the ground floor to the roof only depends on the change in potential energy. That is, the work done is the same no matter what path we took; if we took the stairs or the elevator. Although if we took the elevator, the elevator is doing the work for us. The curves $F(x, y) = C$ are those where no work need be done, such as the heavy box sliding along without accelerating or breaking on a perfectly flat roof, on a cart with incredibly well oiled wheels.

An exact equation is a conservative vector field, and the implicit solution of this equation is the potential function.

1.8.1 Solving exact equations

Now you, the reader, should ask: Where did we solve a differential equation? Well, in applications we generally know M and N , but we do not know F . That is, we may have just

started with $2x + 2y\frac{dy}{dx} = 0$, or perhaps even

$$x + y\frac{dy}{dx} = 0.$$

It is up to us to find some potential F that works. Many different F will work; adding a constant to F does not change the equation. Once we have a potential function F , the equation $F(x, y(x)) = C$ gives an implicit solution of the ODE.

Example 1.8.1: Let us find the general solution to $2x + 2y\frac{dy}{dx} = 0$. Forget we knew what F was.

Solution: If we know that this is an exact equation, we start looking for a potential function F . We have $M = 2x$ and $N = 2y$. If F exists, it must be such that $F_x(x, y) = 2x$. Integrate in the x variable to find

$$F(x, y) = x^2 + A(y), \quad (1.6)$$

for some function $A(y)$. The function A is the “constant of integration”, though it is only constant as far as x is concerned, and may still depend on y . Now differentiate (1.6) in y and set it equal to N , which is what F_y is supposed to be:

$$2y = F_y(x, y) = A'(y).$$

Integrating, we find $A(y) = y^2$. We could add a constant of integration if we wanted to, but there is no need. We found $F(x, y) = x^2 + y^2$. Next for a constant C , we solve

$$F(x, y(x)) = C.$$

for y in terms of x . In this case, we obtain $y = \pm\sqrt{C^2 - x^2}$ as we did before. □

Exercise 1.8.1: Why did we not need to add a constant of integration when integrating $A'(y) = 2y$? Add a constant of integration, say 3, and see what F you get. What is the difference from what we got above, and why does it not matter?

The procedure, once we know that the equation is exact, is:

- (i) Integrate $F_x = M$ in x resulting in $F(x, y) = \text{something} + A(y)$.
- (ii) Differentiate this F in y , and set that equal to N , so that we may find $A(y)$ by integration.

The procedure can also be done by first integrating in y and then differentiating in x . Pretty easy huh? Let's try this again.

Example 1.8.2: Consider now $2x + y + xy\frac{dy}{dx} = 0$.

Solution: OK, so $M = 2x + y$ and $N = xy$. We try to proceed as before. Suppose F exists. Then $F_x(x, y) = 2x + y$. We integrate:

$$F(x, y) = x^2 + xy + A(y)$$

for some function $A(y)$. Differentiate in y and set equal to N :

$$N = xy = F_y(x, y) = x + A'(y).$$

But there is no way to satisfy this requirement! The function xy cannot be written as x plus a function of y . The equation is not exact; no potential function F exists. \square

Is there an easier way to check for the existence of F , other than failing in trying to find it? Turns out there is. Suppose $M = F_x$ and $N = F_y$. Then as long as the second derivatives are continuous,

$$\frac{\partial M}{\partial y} = \frac{\partial^2 F}{\partial y \partial x} = \frac{\partial^2 F}{\partial x \partial y} = \frac{\partial N}{\partial x}.$$

Let us state it as a theorem. Usually this is called the Poincaré Lemma*.

Theorem 1.8.1 (Poincaré)

If M and N are continuously differentiable functions of (x, y) , and $\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$, then near any point there is a function $F(x, y)$ such that $M = \frac{\partial F}{\partial x}$ and $N = \frac{\partial F}{\partial y}$.

The theorem doesn't give us a global F defined everywhere. In general, we can only find the potential locally, near some initial point. By this time, we have come to expect this from differential equations.

Let us return to the example above where $M = 2x + y$ and $N = xy$. Notice $M_y = 1$ and $N_x = y$, which are clearly not equal. The equation is not exact.

Example 1.8.3: Solve

$$\frac{dy}{dx} = \frac{-2x - y}{x - 1}, \quad y(0) = 1.$$

Solution: We write the equation as

$$(2x + y) + (x - 1)\frac{dy}{dx} = 0,$$

so $M = 2x + y$ and $N = x - 1$. Then

$$M_y = 1 = N_x.$$

The equation is exact. Integrating M in x , we find

$$F(x, y) = x^2 + xy + A(y).$$

Differentiating in y and setting to N , we find

$$x - 1 = x + A'(y).$$

*Named for the French polymath Jules Henri Poincaré (1854–1912).

So $A'(y) = -1$, and $A(y) = -y$ will work. Take $F(x, y) = x^2 + xy - y$. We wish to solve $x^2 + xy - y = C$. First let us find C . As $y(0) = 1$ then $F(0, 1) = C$. Therefore $0^2 + 0 \times 1 - 1 = C$, so $C = -1$. Now we solve $x^2 + xy - y = -1$ for y to get

$$y = \frac{-x^2 - 1}{x - 1}.$$

□

Example 1.8.4: Solve

$$-\frac{y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy = 0, \quad y(1) = 2.$$

Solution: We leave to the reader to check that $M_y = N_x$.

This vector field (M, N) is not conservative if considered as a vector field of the entire plane minus the origin. The problem is that if the curve γ is a circle around the origin, say starting at $(1, 0)$ and ending at $(1, 0)$ going counterclockwise, then if F existed we would expect

$$0 = F(1, 0) - F(1, 0) = \int_{\gamma} F_x dx + F_y dy = \int_{\gamma} \frac{-y}{x^2 + y^2} dx + \frac{x}{x^2 + y^2} dy = 2\pi.$$

That is nonsense! We leave the computation of the path integral to the interested reader, or you can consult your multivariable calculus textbook. So there is no potential function F defined everywhere outside the origin $(0, 0)$.

If we think back to the theorem, it does not guarantee such a function anyway. It only guarantees a potential function locally, that is only in some region near the initial point. As $y(1) = 2$ we start at the point $(1, 2)$. Considering $x > 0$ and integrating M in x or N in y , we find

$$F(x, y) = \arctan(y/x).$$

The implicit solution is $\arctan(y/x) = C$. Solving, $y = \tan(C)x$. That is, the solution is a straight line. Solving $y(1) = 2$ gives us that $\tan(C) = 2$, and so $y = 2x$ is the desired solution. See [Figure 1.26](#) on the next page, and note that the solution only exists for $x > 0$. □

Example 1.8.5: Solve

$$x^2 + y^2 + 2y(x+1) \frac{dy}{dx} = 0.$$

Solution: The reader should check that this equation is exact. Let $M = x^2 + y^2$ and $N = 2y(x+1)$. We follow the procedure for exact equations

$$F(x, y) = \frac{1}{3}x^3 + xy^2 + A(y),$$

and

$$2y(x+1) = 2xy + A'(y).$$

Therefore $A'(y) = 2y$ or $A(y) = y^2$ and $F(x, y) = \frac{1}{3}x^3 + xy^2 + y^2$. We try to solve $F(x, y) = C$. We easily solve for y^2 and then just take the square root:

$$y^2 = \frac{C - (1/3)x^3}{x+1}, \quad \text{so} \quad y = \pm \sqrt{\frac{C - (1/3)x^3}{x+1}}.$$

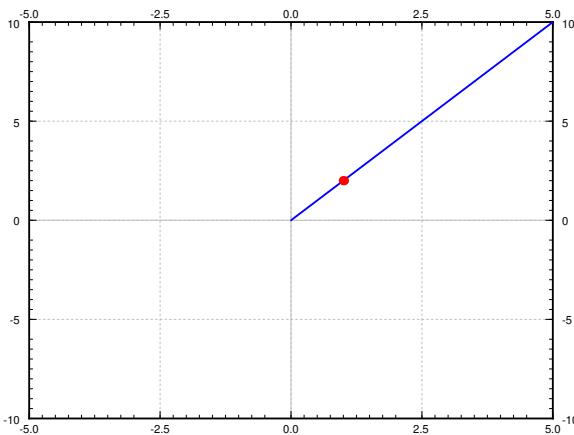


Figure 1.26: Solution to $-\frac{y}{x^2+y^2}dx + \frac{x}{x^2+y^2}dy = 0$, $y(1) = 2$, with initial point marked.

When $x = -1$, the term in front of $\frac{dy}{dx}$ vanishes. You can also see that our solution is not valid in that case. However, one could in that case try to solve for x in terms of y starting from the implicit solution $\frac{1}{3}x^3 + xy^2 + y^2 = C$. The solution is somewhat messy and we leave it as implicit. \square

1.8.2 Integrating factors

Sometimes an equation $M dx + N dy = 0$ is not exact, but it can be made exact by multiplying with a function $u(x, y)$. That is, perhaps for some nonzero function $u(x, y)$,

$$u(x, y)M(x, y)dx + u(x, y)N(x, y)dy = 0$$

is exact. Any solution to this new equation is also a solution to $M dx + N dy = 0$.

In fact, a linear equation

$$\frac{dy}{dx} + p(x)y = f(x), \quad \text{or} \quad (p(x)y - f(x))dx + dy = 0$$

is always such an equation. Let $r(x) = e^{\int p(x)dx}$ be the integrating factor for a linear equation. Multiply the equation by $r(x)$ and write it in the form of $M + N\frac{dy}{dx} = 0$.

$$r(x)p(x)y - r(x)f(x) + r(x)\frac{dy}{dx} = 0.$$

Then $M = r(x)p(x)y - r(x)f(x)$, so $M_y = r(x)p(x)$, while $N = r(x)$, so $N_x = r'(x) = r(x)p(x)$. In other words, we have an exact equation. Integrating factors for linear functions are just a special case of integrating factors for exact equations.

But how do we find the integrating factor u ? Well, given an equation

$$M dx + N dy = 0,$$

u should be a function such that

$$\frac{\partial}{\partial y}[uM] = u_y M + u M_y = \frac{\partial}{\partial x}[uN] = u_x N + u N_x.$$

Therefore,

$$(M_y - N_x)u = u_x N - u_y M.$$

At first it may seem we replaced one differential equation by another. True, but all hope is not lost.

A strategy that often works is to look for a u that is a function of x alone, or a function of y alone. If u is a function of x alone, that is $u(x)$, then we write $u'(x)$ instead of u_x , and u_y is just zero. Then

$$\frac{M_y - N_x}{N}u = u'.$$

In particular, $\frac{M_y - N_x}{N}$ ought to be a function of x alone (not depend on y). If so, then we have a linear equation

$$u' - \frac{M_y - N_x}{N}u = 0.$$

Letting $P(x) = \frac{M_y - N_x}{N}$, we solve using the standard integrating factor method, to find $u(x) = Ce^{\int P(x) dx}$. The constant in the solution is not relevant, we need any nonzero solution, so we take $C = 1$. Then $u(x) = e^{\int P(x) dx}$ is the integrating factor.

Similarly we could try a function of the form $u(y)$. Then

$$\frac{M_y - N_x}{M}u = -u'.$$

In particular, $\frac{M_y - N_x}{M}$ ought to be a function of y alone. If so, then we have a linear equation

$$u' + \frac{M_y - N_x}{M}u = 0.$$

Letting $Q(y) = \frac{M_y - N_x}{M}$, we find $u(y) = Ce^{-\int Q(y) dy}$. We take $C = 1$. So $u(y) = e^{-\int Q(y) dy}$ is the integrating factor.

Example 1.8.6: Solve

$$\frac{x^2 + y^2}{x+1} + 2y \frac{dy}{dx} = 0.$$

Solution: Let $M = \frac{x^2 + y^2}{x+1}$ and $N = 2y$. Compute

$$M_y - N_x = \frac{2y}{x+1} - 0 = \frac{2y}{x+1}.$$

As this is not zero, the equation is not exact. We notice

$$P(x) = \frac{M_y - N_x}{N} = \frac{2y}{x+1} \frac{1}{2y} = \frac{1}{x+1}$$

is a function of x alone. We compute the integrating factor

$$e^{\int P(x) dx} = e^{\ln|x+1|} = |x+1|.$$

Assuming that we want to look at $x > -1$, we multiply our given equation by $(x+1)$ to obtain

$$x^2 + y^2 + 2y(x+1) \frac{dy}{dx} = 0,$$

which is an exact equation that we solved in [Example 1.8.5](#). The solution was

$$y = \pm \sqrt{\frac{C - (1/3)x^3}{x+1}}.$$

If, instead, we had wanted a solution with $x < -1$, we would have needed to multiply by $-(x+1)$, which would have given a very similar result. \square

Example 1.8.7: Solve

$$y^2 + (xy + 1) \frac{dy}{dx} = 0.$$

Solution: First compute

$$M_y - N_x = 2y - y = y.$$

As this is not zero, the equation is not exact. We observe

$$Q(y) = \frac{M_y - N_x}{M} = \frac{y}{y^2} = \frac{1}{y}$$

is a function of y alone. We compute the integrating factor

$$e^{-\int Q(y) dy} = e^{-\ln y} = \frac{1}{y}.$$

Therefore we look at the exact equation

$$y + \frac{xy + 1}{y} \frac{dy}{dx} = 0.$$

The reader should double check that this equation is exact. We follow the procedure for exact equations

$$F(x, y) = xy + A(y),$$

and

$$\frac{xy + 1}{y} = x + \frac{1}{y} = x + A'(y). \quad (1.7)$$

Consequently $A'(y) = \frac{1}{y}$ or $A(y) = \ln y$. Thus $F(x, y) = xy + \ln y$. It is not possible to solve $F(x, y) = C$ for y in terms of elementary functions, so let us be content with the implicit solution:

$$xy + \ln y = C.$$

We are looking for the general solution and we divided by y above. We should check what happens when $y = 0$, as the equation itself makes perfect sense in that case. We plug in $y = 0$ to find the equation is satisfied. So $y = 0$ is also a solution. \square

1.8.3 Exercises

Exercise 1.8.2: Solve the following exact equations, implicit general solutions will suffice:

- | | |
|--|--|
| a) $(2xy + x^2) dx + (x^2 + y^2 + 1) dy = 0$ | b) $x^5 + y^5 \frac{dy}{dx} = 0$ |
| c) $e^x + y^3 + 3xy^2 \frac{dy}{dx} = 0$ | d) $(x + y) \cos(x) + \sin(x) + \sin(x)y' = 0$ |

Exercise 1.8.3:* Solve the following exact equations, implicit general solutions will suffice:

- | | |
|--|--|
| a) $\cos(x) + ye^{xy} + xe^{xy}y' = 0$ | b) $(2x + y) dx + (x - 4y) dy = 0$ |
| c) $e^x + e^y \frac{dy}{dx} = 0$ | d) $(3x^2 + 3y) dx + (3y^2 + 3x) dy = 0$ |

Exercise 1.8.4: Solve the differential equation $(2ye^{2xy} - 2x) + (2xe^{2xy} + \cos(y))y' = 0$

Exercise 1.8.5: Solve the differential equation $(-y \sin(xy) - 2xe^{x^2}) + (-x \sin(xy) + 1)y' = 0$

Exercise 1.8.6: Solve the differential equation $(2x + 3y \sin(xy)) + (3x \sin(xy) - e^y)y' = 0$ with $y(2) = 0$.

Exercise 1.8.7: Solve the differential equation $x + yy' = 0$ with $y(0) = 8$. Write this as an explicit function and determine the interval of x values where the solution is valid.

Exercise 1.8.8: Solve the differential equation $2x - 2 + (8y + 16)y' = 0$ with $y(2) = 0$. Write this as an explicit function and determine the interval of x values where the solution is valid.

Exercise 1.8.9: Find the integrating factor for the following equations making them into exact equations:

- | | |
|---|---|
| a) $e^{xy} dx + \frac{y}{x} e^{xy} dy = 0$ | b) $\frac{e^x + y^3}{y^2} dx + 3x dy = 0$ |
| c) $4(y^2 + x) dx + \frac{2x+2y^2}{y} dy = 0$ | d) $2 \sin(y) dx + x \cos(y) dy = 0$ |

Exercise 1.8.10:* Find the integrating factor for the following equations making them into exact equations:

- | | |
|---|---|
| a) $\frac{1}{y} dx + 3y dy = 0$ | b) $dx - e^{-x-y} dy = 0$ |
| c) $\left(\frac{\cos(x)}{y^2} + \frac{1}{y}\right) dx + \frac{x}{y^2} dy = 0$ | d) $\left(2y + \frac{y^2}{x}\right) dx + (2y + x) dy = 0$ |

Exercise 1.8.11: Suppose you have an equation of the form: $f(x) + g(y) \frac{dy}{dx} = 0$.

- Show it is exact.
- Find the form of the potential function in terms of f and g .

Exercise 1.8.12: Suppose that we have the equation $f(x) dx - dy = 0$.

- Is this equation exact?
- Find the general solution using a definite integral.

Exercise 1.8.13: Find the potential function $F(x, y)$ of the exact equation $\frac{1+xy}{x} dx + \left(\frac{1}{y} + x\right) dy = 0$ in two different ways.

- a) Integrate M in terms of x and then differentiate in y and set to N .
- b) Integrate N in terms of y and then differentiate in x and set to M .

Exercise 1.8.14: A function $u(x, y)$ is said to be a harmonic function if $u_{xx} + u_{yy} = 0$.

- a) Show if u is harmonic, $-u_y dx + u_x dy = 0$ is an exact equation. So there exists (at least locally) the so-called harmonic conjugate function $v(x, y)$ such that $v_x = -u_y$ and $v_y = u_x$.

Verify that the following u are harmonic and find the corresponding harmonic conjugates v :

b) $u = 2xy$ c) $u = e^x \cos y$ d) $u = x^3 - 3xy^2$

Exercise 1.8.15:*

- a) Show that every separable equation $y' = f(x)g(y)$ can be written as an exact equation, and verify that it is indeed exact.
- b) Using this rewrite $y' = xy$ as an exact equation, solve it and verify that the solution is the same as it was in [Example 1.3.1](#).

1.9 Modeling with First Order Equations

Learning Objectives

After this section, you will be able to:

- Write a first-order differential equation to model a physical situation,
- Interpret the solution to a differential equation in the context of a physical problem, and
- Use parameter estimation to approximate physical parameters from data.

One of the main reasons to study and learn about differential equations, particularly for scientists and engineers, is their application and use in mathematical modeling. Since the derivative of a function represents the rate of change of that quantity, if we can use physical or scientific principles to develop an equation for the rate of change of some quantity in terms of the quantity and time, there's a chance that we can write a differential equation for this quantity and solve it to determine how the quantity will change.

1.9.1 Principles of Mathematical Modeling

The process of mathematical modeling involves three main steps. The first of these is to write the model. This part comes from basic science or engineering principles and involves writing a differential equation that fits the given situation. If we can determine the rate at which a quantity will change based on the surrounding factors, we have a good shot of getting to such an equation. One main principle that can be used to write these equations is the accumulation equation, which will be discussed in the next subsection.

The second step of this process is to solve the differential equation. This can mean either an analytic solution or a numeric one, and this is where the work of this class comes into play. We are going through a bunch of different techniques for solving differential equations and analyzing the overall behavior of such equations so that we can use them in this way. The end goal is to get an equation or a graph for how the quantity that we made a model for is going to change in time.

The final step of the process is to validate the model by comparing with experimental data. Once we have written the model and solved the corresponding differential equation, we want to make sure that the model works. To do this, we can take a new version of the original scenario, run the model as well as the physical experiment and see how the results compare. If the results are “close” (in whatever sense makes logical sense for the problem), then we have a good model and can keep it. However, if our results differ significantly, then the model we used probably doesn't apply to this problem. We need to go back to step 1 to try to figure out a better model for the physical situation in order to get more accurate results.

Why do we care about mathematical modeling? The biggest thing that it does from an engineering point of view is reduce the need for repeated testing. If we have a mathematical model that works for a given physical system, we can see how the system will be have under

slightly different conditions and with different initial conditions without needing to run the physical experiment over and over again. We can do all of this testing on the model, and since we have validated the model, we can assume that the actual results will be similar. This also allows us to change some aspects of the physical situation to try to optimize it, but do so just by modifying the mathematical model, not the physical setup. This can significantly cut down on costs and allow for more optimal system design at the same time.

1.9.2 The Accumulation Equation

The accumulation equation is one of the simplest general mathematical formulations that can be used to develop mathematical models. This equation comes down to the fact that the rate of change of some quantity should be equal to the rate at which it is being added minus the rate at which it is being removed. If we let x be the quantity in question, this can be written as

$$\frac{dx}{dt} = \text{rate in} - \text{rate out}. \quad (1.8)$$

This may seem fairly simple. However, it shows up in many places in science and engineering. Any mass or energy balance equations are examples of accumulation equations. These types of equations can also be written for the accumulation of momentum, and doing so for fluids gives rise to the Navier-Stokes equations, providing the basis for several fields of engineering. The examples that we see here will be simpler than that, but the idea is still the same.

Example 1.9.1: A tank initially contains 70 gallons of water and 5 lbs of salt. A solution with salt concentration 0.2 lbs per gallon flows into the tank at a rate of 3 gal/min. The tank is well stirred, and water is removed from the tank at a rate of 3 gal/min. Find the amount of salt in the tank at any time t ? What happens as $t \rightarrow \infty$? Does this make sense?

Solution: To solve this problem, we use the accumulation equation (1.8) on the amount of salt in the tank. In order to compute with this, we recognize that in terms of mass of salt moving into the tank

$$\text{rate in} = \text{flow in} \times \text{concentration in}$$

and similarly for the mass of salt leaving the tank.

If we let x represent the amount of salt in the tank at any time t (which is the goal of the problem), we can write a differential equation for this using the accumulation equation (1.8). This gives us that

$$\frac{dx}{dt} = \text{rate in} - \text{rate out} = \text{flow in} \times \text{concentration in} - \text{flow out} \times \text{concentration out}$$

For this problem, we have that

$$\begin{aligned} \text{flow in} &= 3, \\ \text{concentration in} &= 0.2, \\ \text{flow out} &= 3, \\ \text{concentration out} &= \frac{x}{\text{volume}} = \frac{x}{70}. \end{aligned}$$

The last of these lines comes from the fact that the tank is “well stirred” or “well-mixed.” This implies that the concentration of salt in the water leaving the tank is the same as the concentration in the tank, which we can compute as $\frac{x}{\text{volume}}$. In this case, since the flow rate in and out are both 3 gal/min, the volume of water in the tank is fixed at 70 gallons, so we can put this in the equation.

Therefore, our equation becomes

$$\frac{dx}{dt} = (3 \times 0.2) - \left(3 \times \frac{x}{70}\right).$$

We can rewrite this equation as

$$\frac{dx}{dt} + \frac{3}{70}x = 0.6$$

which we recognize as a first order linear equation. We can then solve this using the method of integrating factors. Our factor $r(t)$ is

$$r(t) = e^{\int p(t) dt} = e^{\int \frac{3}{70} dt} = e^{\frac{3}{70}t},$$

which we can multiply on both sides of the equation to obtain

$$e^{\frac{3}{70}t} \frac{dx}{dt} + e^{\frac{3}{70}t} \frac{3}{70}x = 0.6e^{\frac{3}{70}t}.$$

The left side of this is a product rule derivative, so we can integrate both sides to obtain

$$e^{\frac{3}{70}t}x = 0.6 \frac{70}{3}e^{\frac{3}{70}t} + C.$$

We can then isolate x to get our general solution as

$$x = 14 + Ce^{-\frac{3}{70}t}.$$

Our initial condition tells us that $x(0) = 5$. Plugging this in gives that

$$5 = x(0) = 14 + C \quad \Rightarrow \quad C = -9,$$

so the solution to the initial value problem, and thus our calculation for the amount of salt in the tank at any time t , is

$$x(t) = 14 - 9e^{-\frac{3}{70}t}.$$

As $t \rightarrow \infty$, we see that the exponential term goes to zero. This leaves us with 14 lbs of salt in the tank after a long time. This makes some sense because this would give us a concentration of $\frac{14}{70} = 0.2$ lb/gal, and that was exactly the concentration of the in-flow stream. It makes sense that after a long time of mixing and removing water from the tank, the concentration of the tank would match that of the incoming stream. \square

The same principle works for other types of examples, including those where the volume of the tank is not constant in time.

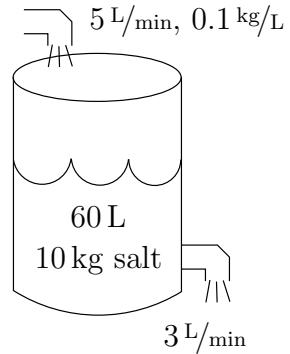
Example 1.9.2: A 100 liter tank contains 10 kilograms of salt dissolved in 60 liters of water. Solution of water and salt (brine) with concentration of 0.1 kilograms per liter is flowing in at the rate of 5 liters a minute. The solution in the tank is well stirred and flows out at a rate of 3 liters a minute. How much salt is in the tank when the tank is full?

Solution: We can again use the accumulation equation to write

$$\frac{dx}{dt} = (\text{flow in} \times \text{concentration in}) - (\text{flow out} \times \text{concentration out}).$$

In this example, we have

$$\begin{aligned} \text{flow in} &= 5, \\ \text{concentration in} &= 0.1, \\ \text{flow out} &= 3, \\ \text{concentration out} &= \frac{x}{\text{volume}} = \frac{x}{60 + (5 - 3)t}. \end{aligned}$$



Our equation is, therefore,

$$\frac{dx}{dt} = (5 \times 0.1) - \left(3 \frac{x}{60 + 2t} \right).$$

Or in the form (1.3)

$$\frac{dx}{dt} + \frac{3}{60 + 2t}x = 0.5.$$

Let us solve. The integrating factor is

$$r(t) = \exp \left(\int \frac{3}{60 + 2t} dt \right) = \exp \left(\frac{3}{2} \ln(60 + 2t) \right) = (60 + 2t)^{3/2}.$$

We multiply both sides of the equation to get

$$\begin{aligned} (60 + 2t)^{3/2} \frac{dx}{dt} + (60 + 2t)^{3/2} \frac{3}{60 + 2t}x &= 0.5(60 + 2t)^{3/2}, \\ \frac{d}{dt} \left[(60 + 2t)^{3/2} x \right] &= 0.5(60 + 2t)^{3/2}, \\ (60 + 2t)^{3/2} x &= \int 0.5(60 + 2t)^{3/2} dt + C, \\ x &= (60 + 2t)^{-3/2} \int \frac{(60 + 2t)^{3/2}}{2} dt + C(60 + 2t)^{-3/2}, \\ x &= (60 + 2t)^{-3/2} \frac{1}{10} (60 + 2t)^{5/2} + C(60 + 2t)^{-3/2}, \\ x &= \frac{60 + 2t}{10} + C(60 + 2t)^{-3/2}. \end{aligned}$$

We need to find C . We know that at $t = 0$, $x = 10$. So

$$10 = x(0) = \frac{60}{10} + C(60)^{-3/2} = 6 + C(60)^{-3/2},$$

or

$$C = 4(60^{3/2}) \approx 1859.03.$$

We are interested in x when the tank is full. The tank is full when $60 + 2t = 100$, or when $t = 20$. So

$$\begin{aligned} x(20) &= \frac{60 + 40}{10} + C(60 + 40)^{-3/2} \\ &\approx 10 + 1859.03(100)^{-3/2} \approx 11.86. \end{aligned}$$

See [Figure 1.27](#) for the graph of x over t .

The concentration when the tank is full is approximately 0.1186 kg/liter , and we started with $\frac{1}{6}$ or 0.167 kg/liter . □

The same ideas apply to problems involving interest compounded continuously. For an interest rate of r , the “rate in,” or the rate at which the money in the account is increasing, is rP where P is the amount of money in the account. Taking this along with other factors that may affect the balance of the account allows us to write a differential equation, which we can solve to determine what the balance will be over time.

Example 1.9.3: A bank account with an interest rate of 6% per year, compounded continuously, starts with a balance of \$30000. The owner of the account withdraws \$50 from the account each month. Find and solve a differential equation for the account balance over time. What is the largest amount that the owner could withdraw each month without the account eventually reaching \$0?

Solution: We will use the function $P(t)$ to model the balance of the account over time, where t is in *years*. Since the owner withdraws \$50 per month, this means that they withdraw \$600 over the course of the year. This means that the differential equation we want is

$$\frac{dP}{dt} = 0.06P - 600 \quad P(0) = 30000.$$

We can solve this equation by the integrating factor method.

$$\begin{aligned} P' - 0.06P &= -600 \\ (e^{-0.06t}P)' &= -600e^{-0.06t} \\ e^{-0.06t}P &= 10000e^{-0.06t} + C \\ P &= 10000 + Ce^{0.06t} \end{aligned}$$

For $P(0) = 30000$, we need to take $C = 20000$. Thus, the solution to the initial value problem is

$$P(t) = 10000 + 20000e^{0.06t}.$$

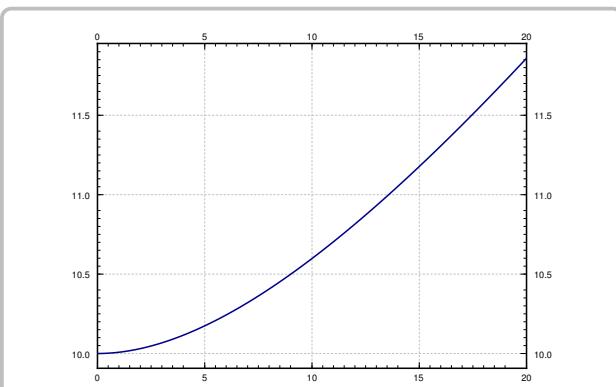


Figure 1.27: Graph of the solution x kilograms of salt in the tank at time t .

Since the coefficient in front of $e^{0.06t}$ is positive, this means that the account balance here will grow in time.

For the second part, we need to adjust the withdrawal amount to see how the solution changes. If we let K be the monthly withdrawal amount, then we have the differential equation

$$\frac{dP}{dt} = 0.06P - 12K \quad P(0) = 30000.$$

The same solution method gives us

$$P(t) = \frac{12K}{0.06} + Ce^{0.06t}.$$

If $C < 0$, then the account balance will eventually go to zero. Therefore, we need $C \geq 0$, and since $P(0) = 30000$, we have that

$$30000 = \frac{12K}{0.06} + C \quad \text{or} \quad C = 30000 - \frac{12K}{0.06}.$$

For this to be equal to zero, we need

$$\frac{12K}{0.06} = 30000 \quad K = 150.$$

Thus, the owner can withdraw \$150 per month and keep the account balance positive. □

1.9.3 Parameter Estimation

One of the most common ways that the mathematical modeling structure can be used to analyze physical problems is the idea of parameter estimation. The idea is that we have a model for how a physical system should behave, but we don't know what the constants should be in the problem. Two main examples of this are Newton's Law of Cooling

$$\frac{dT}{dt} = -k(T - T_s)$$

which models the temperature of an object in an environment of temperature T_s over time, and velocity affected by drag

$$\frac{dv}{dt} = 9.8 - \alpha v^2$$

modeling the velocity of a falling object where the drag force is proportional to the square of the velocity. In both of these cases, the models are well established, but for a given object, we likely do not know the k or α values in the problem. How can we find these values? We can use data from the actual physical problem to try to estimate these parameters.

The easier version of this is to use a single value at a later time to calculate the constant.

Example 1.9.4: An object that obeys Newton's Law of Cooling is placed in an environment at a constant temperature of 20° C. The object starts at 50° C, and after 10 minutes, it has reached a temperature of 40° C. Find a function for the temperature as a function of time.

Solution: Based on Newton's Law of Cooling, we know that the temperature satisfies the differential equation

$$\frac{dT}{dt} = -k(T - T_s) = -k(T - 20)$$

with initial condition $T(0) = 50$, but we do not know the value of k . In order to work this out, we should solve the differential equation with unknown constant k , then figure out which value of k gives us the appropriate temperature after 10 minutes. This is a first order linear equation, which can be rewritten as

$$T' + kT = 20k.$$

The integrating factor we need is e^{kt} , which turns the equation into

$$(e^{kt}T)' = 20ke^{kt}.$$

Integrating both sides and solving for T gives

$$T(t) = 20 + Ce^{-kt}.$$

To satisfy the initial condition, we need that $T(0) = 50$, or $C = 30$. Thus, our solution, still with an unknown constant k , is

$$T(t) = 20 + 30e^{-kt}.$$

To determine the value of k , we need to utilize the other given bit of information: that $T(10) = 40$. Plugging this in gives that

$$40 = 20 + 30e^{-10k}$$

which we can solve for k using logarithms. This will give that

$$\frac{2}{3} = e^{-10k} \quad \Rightarrow \quad k = -\frac{1}{10} \ln \frac{2}{3}.$$

Finally, we can plug that constant into our equation to get the solution for the temperature at any time value,

$$T(t) = 20 + 30e^{-\frac{t}{10} \ln \frac{2}{3}}.$$

□

This method works great if we have the exact measurement from the object at one point in time. However, if the measurements at multiple points in time are known, and if the data is not likely to be exact, then a different method is more applicable. The idea is that we want to minimize the “error” between our predicted result and the physical data that we gather. The method used to minimize the error is the “Least Squared Error” method.

Assume that we want to do this for the drag coefficient problem,

$$\frac{dv}{dt} = 9.8 - \alpha v^2$$

where we do not know, and want to estimate, the value of α . For this method, the data that we gather is a set of velocity values v_1, v_2, \dots, v_n that are obtained at times t_1, t_2, \dots, t_n . For any given value of α , we can solve, either numerically or analytically, the solution v_α to the given differential equation with that value of α . From this solution, we can compute $v_\alpha(t_1), v_\alpha(t_2), \dots, v_\alpha(t_n)$, the value of this solution at each of the times that we gathered data originally. Now, we want to compute the error that we made in choosing this parameter α . This is computed by

$$E(\alpha) = (v_1 - v_\alpha(t_1))^2 + (v_2 - v_\alpha(t_2))^2 + \cdots + (v_n - v_\alpha(t_n))^2$$

which is the sum of the squares of the differences between the gathered data and the predicted solution. In order to find the best possible value of α , we want to minimize this error by choosing different values of α

$$E_{min} = \min_{\alpha} E(\alpha) = \min_{\alpha} \sum_{i=1}^n (v_i - v_\alpha(t_i))^2$$

and whatever value of α gives us this minimum is the optimal choice for that parameter.

The function that we want to minimize here is usually a very complicated function, and we may not even be able to solve the differential equation analytically for any α . Thus, computers are used most often here to solve these types of problems.

Example 1.9.5: An object is falling under the force of gravity, and has a drag component that is proportional to the square of the velocity. Data is gathered on the falling object, and the velocity at a variety of times are given in [Table 1.3](#).

t (s)	v (m/s)
0	0
0.1	0.9797
0.3	2.8625
0.5	4.4750
0.8	6.3828
0.9	6.8360
1.0	7.0334
1.5	8.1612

Table 1.3: Data for estimating drag coefficient using least squared errors.

Use this data to estimate the coefficient of proportionality on the drag term in the equation

$$\frac{dv}{dt} = 9.8 - \alpha v^2.$$

Solution: To solve this problem, we will use the least squared error method implemented in MATLAB. The code we need for this is the following, which makes use of the Optimization Toolbox.

```

global tVals
global vVals

tVals = [0, 0.1, 0.3, 0.5, 0.8, 0.9, 1.0, 1.5];
vVals = [0,0.9797,2.8625,4.4750,6.3828,6.8360,7.0334,8.1612];

[aVal, errVal] = fminbnd(@(a) EstSqError(a), 0, 4)

```

This bit of code inputs the necessary values and uses the `fminbnd` function to find the minimum of the error function on a defined interval. These problems need to be done on a bounded interval, but in most physical situations there is some reasonable window for where the parameter could be. The rest of the code is the definition of the `EstSqError` function.

```

function err = EstSqError(al)

global tVals
global vVals

fun = @(t,v) 9.8 - al.*v.^2;
sol = ode45(fun, [0,3], 0);
vTest = deval(sol, tVals);

err = sum((vVals - vTest).^2)
end

```

The main point of this code is that it takes in a value of α , over which we are trying to minimize, it numerically solves the differential equation with that value of α over a desired range of values, and then compares the inputted `vVals` with the generated `vTest` array, computing the sum of squared errors, and returning the error value.

Running this code results in an α value of 0.1256, with an error of 0.0345. That means that, based on this data, the best approximation to α is 0.1256. □

Note that in the above example, the total error was not zero, and doesn't actually match the coefficient used to generate the data, which was 0.123. This is because noise was added to the data before trying to compute the drag coefficient. In a real world problem, noise would not be added, but a similar effect would arise from slightly inaccurate measurements or round-off errors in the data. While we may not have found the constant exactly, we got really close to it, and could use this as a starting point for further experiments and data validation.

1.9.4 Exercises

Exercise 1.9.1: Suppose there are two lakes located on a stream. Clean water flows into the first lake, then the water from the first lake flows into the second lake, and then water from the second lake flows further downstream. The in and out flow from each lake is 500 liters per hour. The first lake contains 100 thousand liters of water and the second lake

contains 200 thousand liters of water. A truck with 500 kg of toxic substance crashes into the first lake. Assume that the water is being continually mixed perfectly by the stream.

- a) Find the concentration of toxic substance as a function of time in both lakes.
- b) When will the concentration in the first lake be below 0.001 kg per liter?
- c) When will the concentration in the second lake be maximal?

Exercise 1.9.2: Newton's law of cooling states that $\frac{dx}{dt} = -k(x - A)$ where x is the temperature, t is time, A is the ambient temperature, and $k > 0$ is a constant. Suppose that $A = A_0 \cos(\omega t)$ for some constants A_0 and ω . That is, the ambient temperature oscillates (for example night and day temperatures).

- a) Find the general solution.
- b) In the long term, will the initial conditions make much of a difference? Why or why not?

Exercise 1.9.3: Initially 5 grams of salt are dissolved in 20 liters of water. Brine with concentration of salt 2 grams of salt per liter is added at a rate of 3 liters a minute. The tank is mixed well and is drained at 3 liters a minute. How long does the process have to continue until there are 20 grams of salt in the tank?

Exercise 1.9.4: Initially a tank contains 10 liters of pure water. Brine of unknown (but constant) concentration of salt is flowing in at 1 liter per minute. The water is mixed well and drained at 1 liter per minute. In 20 minutes there are 15 grams of salt in the tank. What is the concentration of salt in the incoming brine?

Exercise 1.9.5:* Suppose a water tank is being pumped out at 3 L/min . The water tank starts at 10 L of clean water. Water with toxic substance is flowing into the tank at 2 L/min , with concentration $20t \text{ g/L}$ at time t . When the tank is half empty, how many grams of toxic substance are in the tank (assuming perfect mixing)?

Exercise 1.9.6:* Suppose we have bacteria on a plate and suppose that we are slowly adding a toxic substance such that the rate of growth is slowing down. That is, suppose that $\frac{dP}{dt} = (2 - 0.1t)P$. If $P(0) = 1000$, find the population at $t = 5$.

Exercise 1.9.7:* A cylindrical water tank has water flowing in at I cubic meters per second. Let A be the area of the cross section of the tank in meters. Suppose water is flowing from the bottom of the tank at a rate proportional to the height of the water level. Set up the differential equation for h , the height of the water, introducing and naming constants that you need. You should also give the units for your constants.

Exercise 1.9.8: An object in free fall has a velocity that increases at a rate of 32 ft/s^2 . Due to drag, the velocity decreases at a rate of 0.1 times the velocity of the object squared, when written in feet per second.

- a) Write a differential equation to model the velocity of this object over time.
- b) This equation is autonomous, so draw a phase diagram for this equation and classify all critical points.
- c) What will happen to the velocity if the object is dropped at $t = 0$? What about if the object is thrown downwards at a rate of 10 ft/s ?

Exercise 1.9.9: The number of people in a town that support a given measure decays at a constant rate of 10 people per day. However, the support for the measure can be increased by individuals discussing the issue. This results in an increase of the support at a rate of $ay(1000 - y)$ people per day, where y is the number of people who support the measure, and a is a constant depending on the way in which the issue is being discussed. Write a differential equation to model this situation, and determine the amount of people who will support the measure long-term if a is set to 2.

Exercise 1.9.10: A student has a loan for \$50000 with 5% interest. The student makes \$300 payments on the loan each month.

- a) With this setup, how long does it take the student to pay off the loan? How much money does the student pay over this period of time?
- b) What is the minimal amount the student should pay each month if they want to pay off the loan within 5 years? How much does the student pay over this period?

Exercise 1.9.11: In this exercise, we compare two different young people and their investment strategies. Both of these people are investing in an account with 7.5% annual rate of return. Person 1 invests \$50 a month starting at age 20, and Person 2 invests \$100 per month starting at age 30. Write differential equations to model each of these account balances over time, and compute the amount of money in each account at age 50. Who has more money in the account? Who has invested more money? What would person 2 have to invest each month in order for the two balances to be equal at age 50?

Exercise 1.9.12: Radioactive decay follows similar rules to interest, where a certain portion of the material decays over time, resulting in an equation of the form

$$\frac{dy}{dt} = -ky$$

for some constant k . The half-life of a material is the amount of time that it takes for half of the material to have decayed away. Assume that the half-life of a given substance is T minutes. Find a formula for k , the coefficient in the decay equation, in terms of T .

1.10 Substitution

Attribution: [JL], §1.5.

Learning Objectives

After this section, you will be able to:

- Use substitution to solve more complicated first order equations,
- Use a Bernoulli substitution to solve appropriate first order equations, and
- Use a homogeneity transformation to solve appropriate first order equations.

Just as when solving integrals, one method to try is to change variables to end up with a simpler equation to solve.

1.10.1 Substitution

The equation

$$y' = (x - y + 1)^2$$

is neither separable nor linear. What can we do? How about trying to change variables, so that in the new variables the equation is simpler. We use another variable v , which we treat as a function of x . Let us try

$$v = x - y + 1.$$

We need to figure out y' in terms of v' , v and x . We differentiate (in x) to obtain $v' = 1 - y'$. So $y' = 1 - v'$. We plug this into the equation to get

$$1 - v' = v^2.$$

In other words, $v' = 1 - v^2$. Such an equation we know how to solve by separating variables:

$$\frac{1}{1 - v^2} dv = dx.$$

So

$$\frac{1}{2} \ln \left| \frac{v+1}{v-1} \right| = x + C, \quad \text{or} \quad \left| \frac{v+1}{v-1} \right| = e^{2x+2C}, \quad \text{or} \quad \frac{v+1}{v-1} = De^{2x},$$

for some constant D . Note that $v = 1$ and $v = -1$ are also solutions.

Now we need to “unsubstitute” to obtain

$$\frac{x - y + 2}{x - y} = De^{2x},$$

and also the two solutions $x - y + 1 = 1$ or $y = x$, and $x - y + 1 = -1$ or $y = x + 2$. We solve the first equation for y .

$$\begin{aligned} x - y + 2 &= (x - y)De^{2x}, \\ x - y + 2 &= Dxe^{2x} - yDe^{2x}, \\ -y + yDe^{2x} &= Dxe^{2x} - x - 2, \end{aligned}$$

$$y(-1 + De^{2x}) = Dxe^{2x} - x - 2,$$

$$y = \frac{Dxe^{2x} - x - 2}{De^{2x} - 1}.$$

Note that $D = 0$ gives $y = x + 2$, but no value of D gives the solution $y = x$.

Substitution in differential equations is applied in much the same way that it is applied in calculus. You guess. Several different substitutions might work. There are some general patterns to look for. We summarize a few of these in a table.

When you see	Try substituting
yy'	$v = y^2$
y^2y'	$v = y^3$
$(\cos y)y'$	$v = \sin y$
$(\sin y)y'$	$v = \cos y$
$y'e^y$	$v = e^y$

Usually you try to substitute in the “most complicated” part of the equation with the hopes of simplifying it. The table above is just a rule of thumb. You might have to modify your guesses. If a substitution does not work (it does not make the equation any simpler), try a different one.

1.10.2 Bernoulli equations

There are some forms of equations where there is a general rule for substitution that always works. One such example is the so-called *Bernoulli equation**:

$$y' + p(x)y = q(x)y^n.$$

This equation looks a lot like a linear equation except for the y^n . If $n = 0$ or $n = 1$, then the equation is linear and we can solve it. Otherwise, the substitution $v = y^{1-n}$ transforms the Bernoulli equation into a linear equation. Note that n need not be an integer.

Example 1.10.1: Find the general solution of

$$y' - \frac{4}{3x}y = -\frac{2}{3}y^4$$

Solution: This equation fits the Bernoulli equation structure with $p(x) = -\frac{4}{3x}$ and $q(x) = -\frac{2}{3}$. Since there is a y^4 on the right-hand side, we take $n = 4$ and make the substitution $v = y^{1-4} = y^{-3}$. With this, we see that

$$v' = -3y^{-4}y'$$

*There are several things called Bernoulli equations, this is just one of them. The Bernoullis were a prominent Swiss family of mathematicians. These particular equations are named for **Jacob Bernoulli** (1654–1705).

or $y' = -1/3y^4v'$. Plugging this into the equation gives

$$\begin{aligned} -\frac{1}{3}y^4v' - \frac{4}{3x}y &= -\frac{2}{3}y^4 \\ -\frac{1}{3}v' - \frac{4}{3x}y^{-3} &= -\frac{2}{3} \\ v' + \frac{4}{x}v &= 2 \end{aligned}$$

This last equation is now a first order linear equation, so we can solve it. The integrating factor we are looking for is

$$\mu(x) = e^{\int p(x) dx} = e^{\int \frac{4}{x} dx} = e^{4 \ln x} = x^4,$$

which results in the equation

$$x^4v' + 4x^3v = 2x^4.$$

The left-hand side is $(x^4v)'$, so we can integrate both sides to get

$$x^4v = \frac{2}{5}x^5 + C,$$

or, solving for v ,

$$v(x) = \frac{2}{5}x + \frac{C}{x^4}.$$

However, our original equation was for y , not v . Using the fact that $v = y^{-3}$, we can solve for y as $y = v^{-1/3}$, giving

$$y(x) = \left(\frac{2}{5}x + \frac{C}{x^4} \right)^{-1/3} = \frac{1}{\sqrt[3]{\frac{2}{5}x + \frac{C}{x^4}}}$$

as the general solution to this equation. □

Even if we need to use integrals to write out the solution to these Bernoulli equations, we can still use the substitution method and solve back out for the desired solution at the end.

Example 1.10.2: Solve

$$xy' + y(x+1) + xy^5 = 0, \quad y(1) = 1.$$

Solution: First, the equation is Bernoulli ($p(x) = (x+1)/x$ and $q(x) = -1$). We substitute

$$v = y^{1-5} = y^{-4}, \quad v' = -4y^{-5}y'.$$

In other words, $(-1/4)y^5v' = y'$. So

$$\begin{aligned} xy' + y(x+1) + xy^5 &= 0, \\ \frac{-xy^5}{4}v' + y(x+1) + xy^5 &= 0, \end{aligned}$$

$$\frac{-x}{4}v' + y^{-4}(x+1) + x = 0,$$

$$\frac{-x}{4}v' + v(x+1) + x = 0,$$

and finally

$$v' - \frac{4(x+1)}{x}v = 4.$$

The equation is now linear. We can use the integrating factor method. In particular, we use formula (1.4). Let us assume that $x > 0$ so $|x| = x$. This assumption is OK, as our initial condition is $x = 1$. Let us compute the integrating factor. Here $p(s)$ from formula (1.4) is $\frac{-4(s+1)}{s}$.

$$e^{\int_1^x p(s) ds} = \exp\left(\int_1^x \frac{-4(s+1)}{s} ds\right) = e^{-4x-4 \ln(x)+4} = e^{-4x+4}x^{-4} = \frac{e^{-4x+4}}{x^4},$$

$$e^{-\int_1^x p(s) ds} = e^{4x+4 \ln(x)-4} = e^{4x-4}x^4.$$

We now plug in to (1.4)

$$v(x) = e^{-\int_1^x p(s) ds} \left(\int_1^x e^{\int_1^t p(s) ds} 4 dt + 1 \right)$$

$$= e^{4x-4}x^4 \left(\int_1^x 4 \frac{e^{-4t+4}}{t^4} dt + 1 \right).$$

The integral in this expression is not possible to find in closed form. As we said before, it is perfectly fine to have a definite integral in our solution. Now “unsubstitute”

$$y^{-4} = e^{4x-4}x^4 \left(4 \int_1^x \frac{e^{-4t+4}}{t^4} dt + 1 \right),$$

$$y = \frac{e^{-x+1}}{x \left(4 \int_1^x \frac{e^{-4t+4}}{t^4} dt + 1 \right)^{1/4}}.$$

□

1.10.3 Homogeneous equations

Another type of equations we can solve by substitution are the so-called *homogeneous equations*. Note that this is *not* the same as a homogeneous linear equation. These equations do not have to be linear, and are solved in a very different way. Suppose that we can write the differential equation as

$$y' = F\left(\frac{y}{x}\right).$$

Here we try the substitutions

$$v = \frac{y}{x} \quad \text{and therefore} \quad y' = v + xv'.$$

We note that the equation is transformed into

$$v + xv' = F(v) \quad \text{or} \quad xv' = F(v) - v \quad \text{or} \quad \frac{v'}{F(v) - v} = \frac{1}{x}.$$

Hence an implicit solution is

$$\int \frac{1}{F(v) - v} dv = \ln|x| + C.$$

Example 1.10.3: Solve

$$x^2y' = y^2 + xy, \quad y(1) = 1.$$

Solution: We put the equation into the form $y' = (y/x)^2 + y/x$. We substitute $v = y/x$ to get the separable equation

$$xv' = v^2 + v - v = v^2,$$

which has a solution

$$\begin{aligned} \int \frac{1}{v^2} dv &= \ln|x| + C, \\ \frac{-1}{v} &= \ln|x| + C, \\ v &= \frac{-1}{\ln|x| + C}. \end{aligned}$$

We unsubstitute

$$\begin{aligned} \frac{y}{x} &= \frac{-1}{\ln|x| + C}, \\ y &= \frac{-x}{\ln|x| + C}. \end{aligned}$$

We want $y(1) = 1$, so

$$1 = y(1) = \frac{-1}{\ln|1| + C} = \frac{-1}{C}.$$

Thus $C = -1$ and the solution we are looking for is

$$y = \frac{-x}{\ln|x| - 1}.$$



1.10.4 Exercises

Hint: Answers need not always be in closed form.

Exercise 1.10.1: Solve $y' + y(x^2 - 1) + xy^6 = 0$, with $y(1) = 1$.

Exercise 1.10.2:* Solve $xy' + y + y^2 = 0$, $y(1) = 2$.

Exercise 1.10.3: Solve $2yy' + 1 = y^2 + x$, with $y(0) = 1$.

Exercise 1.10.4:* Solve $xy' + y + x = 0$, $y(1) = 1$.

Exercise 1.10.5: Solve $y' + xy = y^4$, with $y(0) = 1$.

Exercise 1.10.6: Solve $y' + 3y = 2xy^4$.

Exercise 1.10.7: Solve $xy' - 2y = (3x^2 - x^{-3})y^5$ with $y(1) = 2$.

Exercise 1.10.8: Solve $y' + 5y = \frac{e^{2x}}{y^2}$.

Exercise 1.10.9:* Solve $y^2y' = y^3 - 3x$, $y(0) = 2$.

Exercise 1.10.10: Solve $yy' + x = \sqrt{x^2 + y^2}$.

Exercise 1.10.11: Solve $y' = (x + y - 1)^2$.

Exercise 1.10.12: Solve $y' = \frac{x^2 - y^2}{xy}$, with $y(1) = 2$.

Exercise 1.10.13:* Solve $2yy' = e^{y^2 - x^2} + 2x$.

Chapter 2

Higher order linear ODEs

As addressed in [Chapter 1](#), we have a lot of different techniques for solving first order equations. However, not all differential equations are first order. A lot of physical systems in the world operate using higher order equations, particularly second order. Consider the system of a mass hanging from a spring. Newton's second law tells us that the net force on the object equals the mass of the object times its acceleration. However, Hooke's law for springs says that the force the spring exerts on the object is proportional to the distance this object is from the equilibrium position. Therefore, we get a relation between the acceleration of the object and the position. Since the acceleration is the second derivative (in time) of the position of the object, this naturally gives rise to a second order equation.

This means that we want to see what we can do with higher order equations as well. If we can manage to find solutions to these equations as well, then we can address more types of physical problems as well. However, increasing the order of the equation makes it significantly more difficult to find solutions. Even for linear equations, where in first order, we had an explicit method and formula for solutions, we need to put many more restrictions on the equation in order to have a direct method to generate solutions.

2.1 Second order linear ODEs

Attribution: [\[JL\]](#), §2.1.

Learning Objectives

After this section, you will be able to:

- Identify the general second order linear differential equation,
- Determine the characteristic equation for second order constant coefficient equations,
- Find the general solution for constant coefficient equations using exponentials, and
- Determine if two functions are linearly independent.

The general second order ordinary differential equation is of the form

$$y'' = F(x, y, y')$$

for F an arbitrary function of three variables. As with first order equations, if the function F is not in a nice or simple form, there really isn't a hope to find a solution for this. For second order equations, we need to be even more specific about the structure of these equations in order to find solutions than we did for first order.

Definition 2.1.1

The general *second order linear differential equation* is of the form

$$A(x)y'' + B(x)y' + C(x)y = F(x).$$

This equation can be written in *standard form* by dividing through by $A(x)$ to get

$$y'' + p(x)y' + q(x)y = f(x), \quad (2.1)$$

where $p(x) = B(x)/A(x)$, $q(x) = C(x)/A(x)$, and $f(x) = F(x)/A(x)$.

The word *linear* means that the equation contains no powers nor functions of y , y' , and y'' . In the special case when $f(x) = 0$, we have a so-called *homogeneous* equation

$$y'' + p(x)y' + q(x)y = 0. \quad (2.2)$$

We have already seen some second order linear homogeneous equations.

$$y'' + k^2y = 0 \quad \text{Two solutions are: } y_1 = \cos(kx), \quad y_2 = \sin(kx).$$

$$y'' - k^2y = 0 \quad \text{Two solutions are: } y_1 = e^{kx}, \quad y_2 = e^{-kx}.$$

With the examples above, we were able to find solutions. However, notice that these equations don't have functions of x as coefficients of the y term. This means they are constant coefficient equations. It turns out that one of the few ways we can have a guaranteed method for finding solutions to these equations is if they have constant coefficients. For first order, we had a method for every linear equation, but for second order, we only have a formulaic method for constant coefficient and homogeneous linear equations.

If we know two solutions of a linear homogeneous equation, we know many more of them.

Theorem 2.1.1 (Superposition)

Suppose y_1 and y_2 are two solutions of the homogeneous equation (2.2). Then

$$y(x) = C_1y_1(x) + C_2y_2(x),$$

also solves (2.2) for arbitrary constants C_1 and C_2 .

That is, we can add solutions together and multiply them by constants to obtain new and different solutions. We call the expression $C_1y_1 + C_2y_2$ a *linear combination* of y_1 and y_2 .

Let us prove this theorem; the proof is very enlightening and illustrates how linear equations work.

Proof: Let $y = C_1y_1 + C_2y_2$. Then

$$\begin{aligned} y'' + py' + qy &= (C_1y_1 + C_2y_2)'' + p(C_1y_1 + C_2y_2)' + q(C_1y_1 + C_2y_2) \\ &= C_1y_1'' + C_2y_2'' + C_1py_1' + C_2py_2' + C_1qy_1 + C_2qy_2 \\ &= C_1(y_1'' + py_1' + qy_1) + C_2(y_2'' + py_2' + qy_2) \\ &= C_1 \cdot 0 + C_2 \cdot 0 = 0. \quad \square \end{aligned}$$

The proof becomes even simpler to state if we use the operator notation. An *operator* is an object that eats functions and spits out functions (kind of like what a function is, but a function eats numbers and spits out numbers). Define the operator L by

$$L[y] = y'' + py' + qy.$$

The differential equation now becomes $L[y] = 0$. The operator (and the equation) L being *linear* means that $L[C_1y_1 + C_2y_2] = C_1L[y_1] + C_2L[y_2]$. The proof above becomes

$$L[y] = L[C_1y_1 + C_2y_2] = C_1L[y_1] + C_2L[y_2] = C_1 \cdot 0 + C_2 \cdot 0 = 0.$$

Exercise 2.1.1: This fact does not hold if the equation is non-linear. Show that $y_1(t) = e^t$ and $y_2(t) = 1$ solve

$$y'' = \sqrt{y \cdot y'}$$

but $y(t) = e^t + 1$ does not.

Two different solutions to the second equation $y'' - k^2y = 0$ are $y_1 = \cosh(kx)$ and $y_2 = \sinh(kx)$. Let us remind ourselves of the definition, $\cosh x = \frac{e^x + e^{-x}}{2}$ and $\sinh x = \frac{e^x - e^{-x}}{2}$. Therefore, these are solutions by superposition as they are linear combinations of the two exponential solutions.

The functions \sinh and \cosh are sometimes more convenient to use than the exponential. Let us review some of their properties:

$$\begin{array}{ll} \cosh 0 = 1, & \sinh 0 = 0, \\ \frac{d}{dx} [\cosh x] = \sinh x, & \frac{d}{dx} [\sinh x] = \cosh x, \\ \cosh^2 x - \sinh^2 x = 1. & \end{array}$$

Exercise 2.1.2: Derive these properties using the definitions of \sinh and \cosh in terms of exponentials.

2.1.1 Intial Value Problems

For first order equations, a lot of problems were stated as Initial Value Problems, containing both a differential equation and an initial condition of the value of y at some point x_0 . What do these initial condition(s) look like for second order equations?

Example 2.1.1: Solve the second-order differential equation

$$y'' = x.$$

Solution: We can attempt to find a solution to this problem by integrating both sides twice. A first integration gives

$$y' = \frac{x^2}{2} + C$$

and a second integration leads to

$$y = \frac{x^3}{6} + Cx + D$$

for any two constants C and D . We can check that differentiating this y function twice gives us back the function x that we wanted. \square

In the previous example, we ended up with two unknown constants in our answer, whereas for first order equations, we only had one. In order to specify these two constants, we will need to give two additional facts about this function. This could be the value of the function at two points, but more traditionally, it is given as the value of the function y and its first derivative y' at a value x_0 . Fairly often, this value x_0 is 0, but it could be any other number.

Example 2.1.2: Solve the initial value problem

$$y'' = x, \quad y(1) = 2, \quad y'(1) = 3$$

Solution: We previously found our solution with unknown constants as

$$y = \frac{x^3}{6} + Cx + D$$

and also found that

$$y' = \frac{x^2}{2} + C.$$

To find the values of C and D , we need to plug in the two initial conditions into their corresponding functions. The initial value of the derivative gives that

$$3 = y'(1) = \frac{1^2}{2} + C = C + \frac{1}{2}$$

so that we have $C = \frac{5}{2}$. We can then use the initial value of y , along with this C value, to conclude that

$$2 = y(1) = \frac{1^3}{6} + \frac{5}{2}(1) + D = \frac{1}{6} + \frac{5}{2} + D = \frac{16}{6} + D.$$

Solving this out gives that $D = -\frac{4}{6} = -\frac{2}{3}$. Putting these constants in gives that the solution to the initial value problem is

$$y = \frac{x^3}{6} + \frac{5}{2}x - \frac{2}{3}.$$

\square

For first-order equations, we have theorems that told us that solutions existed and were unique, at least on small intervals. Linear first-order equations in particular had a very nice existence and uniqueness theorem (Theorem 1.5.1), guaranteeing existence on a full interval wherever the coefficient functions are continuous. Linear second-order equations have an existence and uniqueness theorem that gives the same type of result when the initial condition is stated properly.

Theorem 2.1.2 (Existence and uniqueness)

Suppose p, q, f are continuous functions on some interval I , a is a number in I , and a, b_0, b_1 are constants. The equation

$$y'' + p(x)y' + q(x)y = f(x),$$

has exactly one solution $y(x)$ defined on the same interval I satisfying the initial conditions

$$y(a) = b_0, \quad y'(a) = b_1.$$

For example, the equation $y'' + k^2y = 0$ with $y(0) = b_0$ and $y'(0) = b_1$ has the solution

$$y(x) = b_0 \cos(kx) + \frac{b_1}{k} \sin(kx).$$

The equation $y'' - k^2y = 0$ with $y(0) = b_0$ and $y'(0) = b_1$ has the solution

$$y(x) = b_0 \cosh(kx) + \frac{b_1}{k} \sinh(kx).$$

Using cosh and sinh in this solution allows us to solve for the initial conditions in a cleaner way than if we have used the exponentials.

2.1.2 Constant Coefficient Equations - Real and Distinct Roots

Now we want to try to solve some of these equations. As discussed earlier in this section, there is no explicit solution method possible for second order equations. However, if we restrict to a very simple case (which is also one that shows up frequently in physical systems) we can start to develop a method for solving these equations. The type of equation we restrict to is linear and constant coefficient equations. *Constant coefficients* means that the functions in front of y'' , y' , and y are constants, they do not depend on x . The most general second order, linear, constant coefficient equation is

$$ay'' + by' + cy = g(x)$$

for real constants a, b, c and an arbitrary function $g(x)$. We will study the solution of nonhomogeneous equations (with $g(x) \neq 0$) in § 2.5. We will first focus on finding general solutions to homogeneous equations, which are of the form

$$ay'' + by' + cy = 0.$$

Consider the problem

$$y'' - 6y' + 8y = 0.$$

This is a second order linear homogeneous equation with constant coefficients, so it fits the type of equation where we want to hunt for solutions. To guess a solution, think of a function that stays essentially the same when we differentiate it, so that we can take the function and its derivatives, add some multiples of these together, and end up with zero. Yes, we are talking about the exponential.

Let us try* a solution of the form $y = e^{rx}$. Then $y' = re^{rx}$ and $y'' = r^2e^{rx}$. Plug in to get

$$\begin{aligned} & y'' - 6y' + 8y = 0, \\ & \underbrace{r^2e^{rx}}_{y''} - 6\underbrace{re^{rx}}_{y'} + 8\underbrace{e^{rx}}_y = 0, \\ & r^2 - 6r + 8 = 0 \quad (\text{divide through by } e^{rx}), \\ & (r - 2)(r - 4) = 0. \end{aligned}$$

Hence, if $r = 2$ or $r = 4$, then e^{rx} is a solution. So let $y_1 = e^{2x}$ and $y_2 = e^{4x}$.

Exercise 2.1.3: Check that y_1 and y_2 are solutions.

So we have found two solutions to this differential equation! That's great, but there may be a few concerning ideas at this point:

- (1) Did we just get lucky with this particular equation?
- (2) How do we know that there aren't other solutions that aren't of the form e^{rx} ? We made that assumption, so we could have missed something.

The second point comes back to the existence and uniqueness theorem. This differential equation satisfies the conditions of the existence and uniqueness theorem. That means that as long as we find *a* solution that can meet any initial condition, then we know that the solution we have found is the *only* solution. We have not yet verified this fact yet (that's coming later), but once we do, we'll know that making this assumption is completely fine, because it got us to a solution that works, and the uniqueness theorem tells us that this is the only solution.

For the first point, let's try to generalize the calculation we did above into a method that will work for more equations. Suppose that we have an equation

$$ay'' + by' + cy = 0, \tag{2.3}$$

where a, b, c are constants. We can take our same assumption that the solution is of the form $y = e^{rx}$ to obtain

$$ar^2e^{rx} + bre^{rx} + ce^{rx} = 0.$$

*Making an educated guess with some parameters to solve for is such a central technique in differential equations, that people sometimes use a fancy name for such a guess: *ansatz*, German for “initial placement of a tool at a work piece.” Yes, the Germans have a word for that.

Divide by e^{rx} to obtain the so-called *characteristic equation* of the ODE:

$$ar^2 + br + c = 0.$$

Solve for the r by using the quadratic formula.

$$r_1, r_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

There are three cases that can arise based on this equation.

- (1) If $b^2 - 4ac > 0$, then we have r_1 and r_2 as two real roots to the equation. This is the same as the example above, and we get $e^{r_1 x}$ and $e^{r_2 x}$ as two solutions. This is the larger class of problems to which this exact process applies.
- (2) If $b^2 - 4ac < 0$, then r_1 and r_2 are complex numbers. We can still use $e^{r_1 x}$ and $e^{r_2 x}$ as solutions, but this runs into some issues, which will be addressed in Section 2.2.
- (3) If $b^2 - 4ac = 0$, then we only get one root, since $r_1 = r_2$. We do get that $e^{r_1 x}$ as a solution, but that's all we get. This is another issue, which is addressed in Section 2.3.

So, as long as we have $b^2 - 4ac > 0$, this method will work to give us two solutions to this differential equation.

Example 2.1.3: Find two values of r so that e^{rx} is a solution to

$$y'' + 3y' - 10y = 0$$

Our first step is to find the characteristic equation by plugging e^{rx} into the equation. This gives that

$$r^2 + 3r - 10 = 0$$

This polynomial factors as $(r - 2)(r + 5)$, so we know that values of $r = 2$ and $r = -5$ will work. This means (check this!) that e^{2x} and e^{-5x} solve this differential equation.

2.1.3 Linear Independence

Since e^{2x} and e^{-5x} solve the linear differential equation in the previous example, we know that superposition applies, so that $C_1 e^{2x} + C_2 e^{-5x}$ solves the differential equation for any C_1 and C_2 . The last thing to check is that we can pick C_1 and C_2 in order to meet any initial condition that we want. If this is possible, then we know that our method using the characteristic equation to find e^{2x} and e^{-5x} as solutions was enough to always solve this problem.

Let's work this out. Assume that we are given b_0 and b_1 and want to solve the initial value problem

$$y'' + 3y' - 10y = 0 \quad y(0) = b_0, \quad y'(0) = b_1.$$

We want to do this by picking C_1 and C_2 in the expression $y = C_1 e^{2x} + C_2 e^{-5x}$. Since

$$y' = 2C_1 e^{2x} - 5C_2 e^{-5x}$$

we can plug zero into this equation and the equation for y to get that we would need to have

$$\begin{aligned} b_0 &= y(0) = C_1 + C_2 \\ b_1 &= y'(0) = 2C_1 - 5C_2. \end{aligned}$$

We can solve this system of equations by elimination. Multiplying the first equation by 5 adding them together gives

$$5b_0 + b_1 = 7C_1$$

so that

$$C_1 = \frac{5b_0 + b_1}{7}.$$

We can then compute the value of C_2 as

$$C_2 = b_0 - C_1 = b_0 - \frac{5b_0 + b_1}{7} = \frac{2b_0 - b_1}{7}.$$

Therefore, we can appropriate values of C_1 and C_2 that will meet the initial conditions for arbitrary values b_0 and b_1 . This is great! This means that our method of finding solutions was sufficient for this problem.

Let's look at this situation in more generality. Assume that we have two solutions y_1 and y_2 that solve a second order linear, homogeneous differential equation, and we want to know if $C_1y_1 + C_2y_2$ can meet any initial condition for this problem. We have two unknowns and two equations ($y(x_0)$ and $y'(x_0)$ for some value x_0), so it should work out.

We can carry out the same steps as above. If we have initial conditions $y(x_0) = b_0$ and $y'(x_0) = b_1$, we want to satisfy

$$\begin{aligned} b_0 &= y(x_0) = C_1y_1(x_0) + C_2y_2(x_0) \\ b_1 &= y'(x_0) = C_1y'_1(x_0) + C_2y'_2(x_0), \end{aligned}$$

which we get by taking the derivative of $y(x) = C_1y_1(x) + C_2y_2(x)$ and plugging in x_0 . We will again use elimination to solve this. We can multiply the first equation by $y'_1(x_0)$, multiply the second by $y_1(x_0)$, and subtract them. This will cancel out the C_1 term, leaving us with

$$b_0y'_1(x_0) - b_1y_1(x_0) = C_2(y'_1(x_0)y_2(x_0) - y_1(x_0)y'_2(x_0)).$$

We want to solve for C_2 here, and once we do that, solving for C_1 happens by plugging back into one of the original equations. Most of the time, this will be completely fine, but there's one issue left. We can't divide by zero. So to be able to solve these equations for C_1 and C_2 , we need to know that

$$y'_1(x_0)y_2(x_0) - y_1(x_0)y'_2(x_0) \neq 0 \tag{2.4}$$

This relation tells us that the two solutions y_1 and y_2 are different enough to allow us to meet every initial condition for the differential equation. This condition is so important to the study of second order linear equations that we give it a name. We say that two solutions y_1 and y_2 are *linearly independent* if the only way to make the expression

$$c_1y_1 + c_2y_2 = 0$$

is by setting both $c_1 = 0$ and $c_2 = 0$. If there are such constants, we can also rearrange the equation to give

$$y_1 = -\frac{c_2}{c_1}y_2$$

which says that y_1 is a constant multiple of y_2 . Thus, if we have y_1 and y_2 , and there is no constant A so that $y_1 = Ay_2$, then these functions are linearly independent. For two solutions of a differential equation (which is more specific than just having two random functions), two solutions being linearly independent is equivalent to 2.4 holding for any* value x_0 where they are defined. Our work and calculations above leads to the following theorem:

Theorem 2.1.3

Let p, q be continuous functions. Let y_1 and y_2 be two linearly independent solutions to the homogeneous equation (2.2). Then every other solution is of the form

$$y = C_1y_1 + C_2y_2$$

for some constants C_1 and C_2 . That is, $y = C_1y_1 + C_2y_2$ is the general solution.

Note that this theorem works for all linear homogeneous equations, not just constant coefficients ones. For example, we found the solutions $y_1 = \sin x$ and $y_2 = \cos x$ for the equation $y'' + y = 0$. It is not hard to see that sine and cosine are not constant multiples of each other. If $\sin x = A \cos x$ for some constant A , we let $x = 0$ and this would imply $A = 0$. But then $\sin x = 0$ for all x , which is preposterous. So y_1 and y_2 are linearly independent. We could also have checked this by taking derivatives and plugging in zero. Since

$$y_1(0) = 0 \quad y'_1(0) = 1 \quad y_2(0) = 1 \quad y'_2(0) = 0$$

we have that

$$y'_1(0)y_2(0) - y_1(0)y'_2(0) = (1)(1) - (0)(0) = 1 \neq 0$$

so these solutions are linearly independent. Hence,

$$y = C_1 \cos x + C_2 \sin x$$

is the general solution to $y'' + y = 0$.

For two functions, checking linear independence is rather simple. Let us see another example. Consider $y'' - 2x^{-2}y = 0$. Then $y_1 = x^2$ and $y_2 = 1/x$ are solutions. To see that they are linearly independent, suppose one is a multiple of the other: $y_1 = Ay_2$, we just have to find out that A cannot be a constant. In this case we have $A = y_1/y_2 = x^3$, this most decidedly not a constant. So $y = C_1x^2 + C_21/x$ is the general solution.

Now, back to our discussion of constant coefficient equations. If $b^2 - 4ac > 0$, then we have two distinct real roots r_1 and r_2 , giving rise to solutions of the form $y_1(x) = e^{r_1 x}$ and $y_2(x) = e^{r_2 x}$. Using condition 2.4 with $x_0 = 0$, we compute

$$y'_1(0)y_2(0) - y_1(0)y'_2(0) = (r_1)(1) - (1)(r_2) = r_1 - r_2.$$

*Abel's Theorem, another theoretical result, says that this function $y'_1y_2 - y_1y'_2$ is either always zero or never zero. That means that any one value can be checked to determine if two solutions are linearly independent. Picking 0 is usually a convenient choice.

Since $r_1 \neq r_2$, this expression is not zero, so the two solutions are linearly independent. Therefore, in this case, we know that the general solution will be

$$y = C_1 e^{r_1 x} + C_2 e^{r_2 x}.$$

Example 2.1.4: Solve the initial value problem

$$y'' - 2y' - y = 0 \quad y(0) = 2, \quad y'(0) = 3.$$

Solution: We start by looking for the characteristic equation of this differential equation and finding its roots. The characteristic equation is

$$r^2 - 2r - 1 = 0$$

which has roots

$$r = \frac{2 \pm \sqrt{(-2)^2 - 4(1)(-1)}}{2} = \frac{2 \pm \sqrt{8}}{2} = 1 \pm \sqrt{2}.$$

There are two real and distinct roots, so we know that the two solutions $y_1(x) = e^{(1+\sqrt{2})x}$ and $y_2(x) = e^{(1-\sqrt{2})x}$ are linearly independent, so we have that the general solution to this problem is

$$y(x) = C_1 e^{(1+\sqrt{2})x} + C_2 e^{(1-\sqrt{2})x}.$$

Next, we need to find the constants C_1 and C_2 to meet the initial conditions. We can see that, by computing the first derivative,

$$\begin{aligned} y(x) &= C_1 e^{(1+\sqrt{2})x} + C_2 e^{(1-\sqrt{2})x}, \\ y'(x) &= (1 + \sqrt{2})C_1 e^{(1+\sqrt{2})x} + (1 - \sqrt{2})C_2 e^{(1-\sqrt{2})x}, \end{aligned}$$

and plugging in $x = 0$ gives that we want C_1 and C_2 to solve

$$\begin{aligned} 2 &= C_1 + C_2, \\ 3 &= (1 + \sqrt{2})C_1 + (1 - \sqrt{2})C_2. \end{aligned}$$

We can solve this by any method. One trick at the start is to subtract equation 1 from equation 2, giving that

$$\begin{aligned} 2 &= C_1 + C_2, \\ 1 &= \sqrt{2}C_1 - \sqrt{2}C_2, \end{aligned}$$

which can be rewritten as

$$\begin{aligned} 2 &= C_1 + C_2, \\ \frac{1}{\sqrt{2}} &= C_1 - C_2. \end{aligned}$$

Adding these equations together and dividing by 2 gives that

$$2C_1 = 2 + \frac{1}{\sqrt{2}}$$

so that $C_1 = 1 + \frac{1}{2\sqrt{2}}$, and since $C_1 + C_2 = 2$, we have that $C_2 = 1 - \frac{1}{2\sqrt{2}}$. Therefore, the solution to the desired initial value problem is

$$y(x) = \left(1 + \frac{1}{2\sqrt{2}}\right) e^{(1+\sqrt{2})x} + \left(1 - \frac{1}{2\sqrt{2}}\right) e^{(1-\sqrt{2})x}. \quad \square$$

2.1.4 Exercises

Exercise 2.1.4: Show that $y = e^x$ and $y = e^{2x}$ are linearly independent.

Exercise 2.1.5:* Are $\sin(x)$ and e^x linearly independent? Justify.

Exercise 2.1.6:* Are e^x and e^{x+2} linearly independent? Justify.

Exercise 2.1.7:* Guess a solution to $y'' + y' + y = 5$.

Exercise 2.1.8: Take $y'' + 5y = 10x + 5$. Find (guess!) a solution.

Exercise 2.1.9: Verify that $y_1(t) = e^t \cos(2t)$ and $y_2(t) = e^t \sin(2t)$ both solve $y'' - 2y' + 5y = 0$. Are these two solutions linearly independent? What does that mean about the general solution to $y'' - 2y' + 5y = 0$?

Exercise 2.1.10: Prove the superposition principle for nonhomogeneous equations. Suppose that y_1 is a solution to $Ly_1 = f(x)$ and y_2 is a solution to $Ly_2 = g(x)$ (same linear operator L). Show that $y = y_1 + y_2$ solves $Ly = f(x) + g(x)$.

Exercise 2.1.11: Determine the maximal interval of existence of the solution to the differential equation

$$(t-5)y'' + \frac{1}{t+1}y' + e^t y = \frac{\cos(t)}{t^2+1}$$

with initial condition $y(3) = 8$. What about if the initial condition is $y(-3) = 4$?

Exercise 2.1.12: For the equation $x^2y'' - xy' = 0$, find two solutions, show that they are linearly independent and find the general solution. Hint: Try $y = x^r$.

Exercise 2.1.13:* Find the general solution to $xy'' + y' = 0$. Hint: It is a first order ODE in y' .

Exercise 2.1.14: Find the general solution of $2y'' + 2y' - 4y = 0$.

Exercise 2.1.15: Solve $y'' + 9y' = 0$ with $y(0) = 1$, $y'(0) = 1$.

Exercise 2.1.16: Find the general solution of $y'' + 9y' - 10y = 0$.

Exercise 2.1.17: Find the general solution to $y'' - 3y' - 4y = 0$.

Exercise 2.1.18: Find the general solution to $y'' + 6y' + 8y = 0$.

Exercise 2.1.19: Find the solution to $y'' - 3y' + 2y = 0$ with $y(0) = 3$ and $y'(0) = -1$.

Exercise 2.1.20: Find the solution to $y'' + y' - 12y = 0$ with $y(0) = 1$ and $y'(0) = -2$.

Exercise 2.1.21:* Find the general solution to $y'' + 4y' + 2y = 0$.

Exercise 2.1.22:* Find the solution to $2y'' + y' - 3y = 0$, $y(0) = a$, $y'(0) = b$.

Exercise 2.1.23:* Find the solution to $y'' - (\alpha + \beta)y' + \alpha\beta y = 0$, $y(0) = a$, $y'(0) = b$, where α , β , a , and b are real numbers, and $\alpha \neq \beta$.

Exercise 2.1.24:* Construct an equation such that $y = C_1 e^{3x} + C_2 e^{-2x}$ is the general solution.

Exercise 2.1.25:* Write down an equation (guess) for which we have the solutions e^x and e^{2x} . Hint: Try an equation of the form $y'' + Ay' + By = 0$ for constants A and B , plug in both e^x and e^{2x} and solve for A and B .

Equations of the form $ax^2y'' + bxy' + cy = 0$ are called *Euler's equations* or *Cauchy-Euler equations*. They are solved by trying $y = x^r$ and solving for r (assume that $x \geq 0$ for simplicity).

Exercise 2.1.26: Suppose that $(b - a)^2 - 4ac > 0$.

- a) Find a formula for the general solution of $ax^2y'' + bxy' + cy = 0$. Hint: Try $y = x^r$ and find a formula for r .
- b) What happens when $(b - a)^2 - 4ac = 0$ or $(b - a)^2 - 4ac < 0$?

We will revisit the case when $(b - a)^2 - 4ac < 0$ later.

Exercise 2.1.27: Same equation as in [Exercise 2.1.26](#). Suppose $(b - a)^2 - 4ac = 0$. Find a formula for the general solution of $ax^2y'' + bxy' + cy = 0$. Hint: Try $y = x^r \ln x$ for the second solution.

2.2 Complex Roots and Euler's Formula

Attribution: [JL], §2.2.

Learning Objectives

After this section, you will be able to:

- Understand the basics of complex numbers,
- Use complex numbers to find complex solutions to second order constant coefficient equations, and
- Use Euler's formula to find real-valued general solutions to these second order equations.

The next case to consider for constant coefficient second order equations is the one where $b^2 - 4ac < 0$. This results in two roots r_1 and r_2 , but they are complex roots. In order to solve differential equations with $b^2 - 4ac < 0$, we need to be able to manipulate and use some properties of complex numbers. Complex numbers may seem a strange concept, especially because of the terminology. There is nothing imaginary or really complicated about complex numbers. For more background information on complex numbers, see [Appendix B.2](#).

To start with, we define $i = \sqrt{-1}$. Since this is the square root of a negative number, this i is not a real number. A complex number is written in the form $z = x + iy$ where x and y are real numbers. For a complex number $x + iy$ we call x the *real part* and y the *imaginary part* of the number. Often the following notation is used,

$$\operatorname{Re}(x + iy) = x \quad \text{and} \quad \operatorname{Im}(x + iy) = y.$$

The real numbers are contained in the complex numbers if we have the imaginary part being zero.

When trying to do arithmetic with complex numbers, we treat i as though it is a variable, and do computations just as we would with polynomials. The important fact that we will use to simplify is the fact that since $i = \sqrt{-1}$, we have that $i^2 = -1$. So whenever we see i^2 , we replace it by -1 . For example,

$$(2 + 3i)(4i) - 5i = (2 \times 4)i + (3 \times 4)i^2 - 5i = 8i + 12(-1) - 5i = -12 + 3i.$$

The numbers i and $-i$ are the two roots of $r^2 + 1 = 0$. Engineers often use the letter j instead of i for the square root of -1 . We use the mathematicians' convention and use i .

Exercise 2.2.1: Make sure you understand (that you can justify) the following identities:

- | | |
|--|--|
| a) $i^2 = -1, i^3 = -i, i^4 = 1,$ | b) $\frac{1}{i} = -i,$ |
| c) $(3 - 7i)(-2 - 9i) = \dots = -69 - 13i,$ | d) $(3-2i)(3+2i) = 3^2 - (2i)^2 = 3^2 + 2^2 = 13,$ |
| e) $\frac{1}{3-2i} = \frac{1}{3-2i} \frac{3+2i}{3+2i} = \frac{3+2i}{13} = \frac{3}{13} + \frac{2}{13}i.$ | |

In order to solve differential equations where the characteristic equation has complex roots, we need to deal with exponential e^{a+bi} of complex numbers. We do this by writing down the Taylor series and plugging in the complex number. Because most properties of the exponential can be proved by looking at the Taylor series, these properties still hold for the complex exponential. For example the very important property: $e^{x+y} = e^x e^y$. This means that $e^{a+ib} = e^a e^{ib}$. Hence if we can compute e^{ib} , we can compute e^{a+ib} . For e^{ib} we use the so-called *Euler's formula*.

Theorem 2.2.1 (Euler's formula)

$$e^{i\theta} = \cos \theta + i \sin \theta \quad \text{and} \quad e^{-i\theta} = \cos \theta - i \sin \theta.$$

In other words, $e^{a+ib} = e^a (\cos(b) + i \sin(b)) = e^a \cos(b) + ie^a \sin(b)$.

Exercise 2.2.2: Using Euler's formula, check the identities:

$$\cos \theta = \frac{e^{i\theta} + e^{-i\theta}}{2} \quad \text{and} \quad \sin \theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}.$$

Exercise 2.2.3: Double angle identities: Start with $e^{i(2\theta)} = (e^{i\theta})^2$. Use Euler on each side and deduce:

$$\cos(2\theta) = \cos^2 \theta - \sin^2 \theta \quad \text{and} \quad \sin(2\theta) = 2 \sin \theta \cos \theta.$$

2.2.1 Complex roots

Suppose the equation $ay'' + by' + cy = 0$ has the characteristic equation $ar^2 + br + c = 0$ that has complex roots. By the quadratic formula, the roots are $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$. These roots are complex if $b^2 - 4ac < 0$. In this case the roots are

$$r_1, r_2 = \frac{-b}{2a} \pm i \frac{\sqrt{4ac - b^2}}{2a}.$$

As you can see, we always get a pair of roots of the form $\alpha \pm i\beta$. In this case we can still write the solution as

$$y = C_1 e^{(\alpha+i\beta)x} + C_2 e^{(\alpha-i\beta)x}.$$

However, the exponential is now complex-valued, and so (real) linear combinations of these solutions will be complex valued. If we are using these equations to model physical problems, the answer should be real-valued, as the position of a mass-on-a-spring can not be a complex number. To do this, we need to determine two real-valued, linearly independent solutions to this differential equation.

To do this, we use the following result.

Theorem 2.2.2

Consider the differential equation

$$y'' + p(x)y' + q(x)y = 0$$

where $p(t)$ and $q(t)$ are *real-valued* continuous functions on some interval I . If y is a complex-valued solution to this differential equation and we can split $y(x) = u(x) + iv(x)$ into its real and imaginary parts u and v , then u and v are both solutions to $y'' + p(x)y' + q(x)y = 0$.

Proof. This is based on the fact that the differential equation is linear. We can compute derivatives of y

$$\begin{aligned} y(x) &= u(x) + iv(x) \\ y'(x) &= u'(x) + iv'(x) . \\ y''(x) &= u''(x) + iv''(x) \end{aligned}$$

Then, we can plug this into the differential equation

$$\begin{aligned} 0 &= y'' + p(x)y' + q(x)y \\ &= u''(x) + iv''(x) + p(x)(u'(x) + iv'(x)) + q(x)(u(x) + iv(x)) . \\ 0 &= u''(x) + p(x)u'(x) + q(x)u(x) + i(v''(x) + p(x)v'(x) + q(x)v(x)) \end{aligned}$$

Since the equation at the end of this chain is equal to zero, it must be zero as a complex number, which means that both the real and imaginary parts must be zero. This means that

$$\begin{aligned} u''(x) + p(x)u'(x) + q(x)u(x) &= 0 \\ v''(x) + p(x)v'(x) + q(x)v(x) &= 0 \end{aligned}$$

so that both u and v solve the original differential equation. \square

To use this to solve the problem at hand, we have our solution

$$y_1(x) = e^{\alpha+i\beta x}$$

and we need to split this into its real and imaginary parts. Since

$$y_1 = e^{\alpha x} \cos(\beta x) + ie^{\alpha x} \sin(\beta x),$$

the real and imaginary parts of this function are

$$\begin{aligned} u(x) &= e^{\alpha x} \cos(\beta x) \\ v(x) &= e^{\alpha x} \sin(\beta x) \end{aligned}$$

which, by the previous theorem, we know are also solutions. These are two solutions to our original differential equation that are also real-valued!

Exercise 2.2.4: For $\beta \neq 0$, check that $e^{\alpha x} \cos(\beta x)$ and $e^{\alpha x} \sin(\beta x)$ are linearly independent.

With that fact, we have the following theorem.

Theorem 2.2.3

Take the equation

$$ay'' + by' + cy = 0.$$

If the characteristic equation has the roots $\alpha \pm i\beta$ (when $b^2 - 4ac < 0$), then the general solution is

$$y = C_1 e^{\alpha x} \cos(\beta x) + C_2 e^{\alpha x} \sin(\beta x).$$

Example 2.2.1: Find the general solution of $y'' + k^2 y = 0$, for a constant $k > 0$.

Solution: The characteristic equation is $r^2 + k^2 = 0$. Therefore, the roots are $r = \pm ik$, and by the theorem, we have the general solution

$$y = C_1 \cos(kx) + C_2 \sin(kx). \quad \square$$

Example 2.2.2: Find the solution of $y'' - 6y' + 13y = 0$, $y(0) = 0$, $y'(0) = 10$.

Solution: The characteristic equation is $r^2 - 6r + 13 = 0$. By completing the square we get $(r - 3)^2 + 2^2 = 0$ and hence the roots are $r = 3 \pm 2i$. By the theorem we have the general solution

$$y = C_1 e^{3x} \cos(2x) + C_2 e^{3x} \sin(2x).$$

To find the solution satisfying the initial conditions, we first plug in zero to get

$$0 = y(0) = C_1 e^0 \cos 0 + C_2 e^0 \sin 0 = C_1.$$

Hence, $C_1 = 0$ and $y = C_2 e^{3x} \sin(2x)$. We differentiate,

$$y' = 3C_2 e^{3x} \sin(2x) + 2C_2 e^{3x} \cos(2x).$$

We again plug in the initial condition and obtain $10 = y'(0) = 2C_2$, or $C_2 = 5$. The solution we are seeking is

$$y = 5e^{3x} \sin(2x). \quad \square$$

In this previous example, we can get a fairly good idea of how to sketch out the graph of this function. Since $\sin(2x)$ oscillates between -1 and 1 , the graph of $y = 5e^{3x} \sin(2x)$ will oscillate between the graphs of $5e^{3x}$ and $-5e^{3x}$. These curves that surround the graph of the solution are called *envelope curves* for the solution. In Figure 2.1, this phenomenon is illustrated for the function $y = 2e^x \sin(5x)$.

This is simple when there is only one term in the function we want to draw. When both sine and cosine terms appear, this can get more tricky, but we can still work it out. The solution will look something like

$$y = Ae^{\alpha x} \cos(\beta x) + Be^{\alpha x} \sin(\beta x).$$

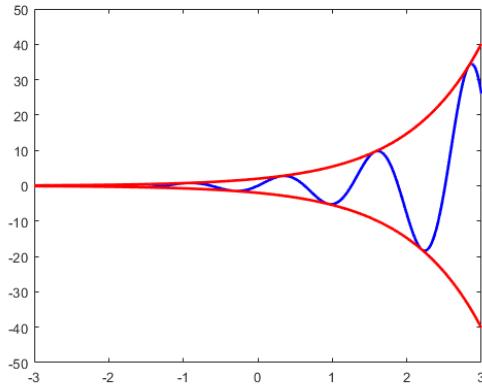


Figure 2.1: Plot of the function $y = 2e^x \sin(5x)$ with envelope curves.

We can first factor out an $e^{\alpha x}$, and then we want to write $A \cos(\beta x) + B \sin(\beta x)$ as a single trigonometric function. The identity we want to use here is the trigonometric identity

$$\cos(\beta x - \delta) = \cos(\delta) \cos(\beta x) + \sin(\delta) \sin(\beta x).$$

If there is an angle δ so that $A = \cos(\delta)$ and $B = \sin(\delta)$, then we could write

$$A \cos(\beta x) + B \sin(\beta x) = \cos(\beta x - \delta)$$

and we would be done. However, this does not always happen; the main issue being that $\cos^2(\delta) + \sin^2(\delta) = 1$ for all δ , but it is not necessarily the case that $A^2 + B^2 = 1$. But we can force this last condition. If we define $R = \sqrt{A^2 + B^2}$, then we can rewrite this expression as

$$\begin{aligned} A \cos(\beta x) + B \sin(\beta x) &= R \left(\frac{A}{\sqrt{A^2 + B^2}} \cos(\beta x) + \frac{B}{\sqrt{A^2 + B^2}} \sin(\beta x) \right) \\ &= R (\cos(\delta) \cos(\beta x) + \sin(\delta) \sin(\beta x)) \\ &= R \cos(\beta x - \delta) \end{aligned}$$

where δ is the angle so that

$$\cos(\delta) = \frac{A}{R} \quad \sin(\delta) = \frac{B}{R}$$

and such an angle will always exist. Therefore, we can represent the original solution

$$y = Ae^{\alpha x} \cos(\beta x) + Be^{\alpha x} \sin(\beta x)$$

as

$$y = Re^{\alpha x} \cos(\beta x - \delta)$$

where

$$R = \sqrt{A^2 + B^2} \quad \cos(\delta) = \frac{A}{R} \quad \sin(\delta) = \frac{B}{R}.$$

Therefore, the envelope curves for this solution will be

$$y = \pm Re^{\alpha x}.$$

Example 2.2.3: Find the solution to the initial value problem

$$y'' + 2y' + 5y = 0 \quad y(0) = 1, \quad y'(0) = 5.$$

Determine a value T where the solution $y(x)$ satisfies $|y(x)| < 0.1$ for all $x > T$.

Solution: We solve the initial value problem by normal techniques from this section. The characteristic equation is $r^2 + 2r + 5 = 0$, which has roots $r = -1 \pm 2i$. Therefore, the general solution of the differential equation is

$$y = C_1 e^{-x} \cos(2x) + C_2 e^{-x} \sin(2x).$$

Plugging in 0 gives that $y(0) = 1 = C_1$, and the derivative of this general solution is

$$y' = -C_1 e^{-x} \cos(2x) - 2C_1 e^{-x} \sin(2x) - C_2 e^{-x} \sin(2x) + 2C_2 e^{-x} \cos(2x).$$

Plugging in 0 here gives

$$y'(0) = -C_1 + 2C_2.$$

Since $C_1 = 1$, this gives that $C_2 = 3$. So, our solution is

$$y(x) = e^{-x} \cos(2x) + 3e^{-x} \sin(2x).$$

Through the work above, we can find $R = \sqrt{1+9} = \sqrt{10}$. Therefore, the envelope curves for the solution are

$$\pm\sqrt{10}e^{-x}.$$

In order to find this threshold T where the solution will stay within 0.1 of zero, we need to figure out when this envelope curve gets in that range. We can solve

$$0.1 = \sqrt{10}e^{-T} \quad T = -\ln\left(\frac{0.1}{\sqrt{10}}\right) \approx 3.454.$$

So, for all values of x larger than 3.454, the solution will be within 0.1 of zero. This is illustrated in [Figure 2.2](#). Note that we did not find the *best* value T here, as it probably could be made smaller using the actual solution. But we did find a value of T that works. [|](#)

2.2.2 Exercises

Exercise 2.2.5:* Write $3\cos(2x) + 3\sin(2x)$ in the form $R\cos(\beta x - \delta)$.

Exercise 2.2.6: Write $2\cos(3x) + \sin(3x)$ in the form $R\cos(\beta x - \delta)$.

Exercise 2.2.7: Write $3\cos(x) - 4\sin(x)$ in the form $R\cos(\beta x - \delta)$.

Exercise 2.2.8: Show that $e^{2x}\cos(x)$ and $e^{2x}\sin(x)$ are linearly independent.

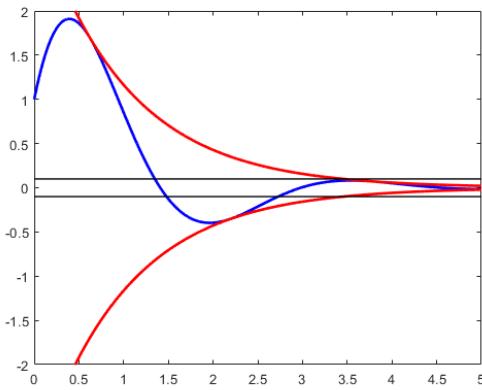


Figure 2.2: Plot of the function $e^{-x} \cos(2x) + 3e^{-x} \sin(2x)$ with envelope curves illustrating the bounds on the function for large values of x .

Exercise 2.2.9: Find the general solution of $2y'' + 50y = 0$.

Exercise 2.2.10: Find the general solution of $y'' - 6y' + 13y = 0$.

Exercise 2.2.11: Find the solution to $y'' - 2y' + 5y = 0$ with $y(0) = 3$ and $y'(0) = 2$.

Exercise 2.2.12: Find the general solution of $y'' + 2y' - 3y = 0$.

Exercise 2.2.13:* Find the solution to $2y'' + y' + y = 0$, $y(0) = 1$, $y'(0) = -2$.

Exercise 2.2.14:* Find the solution to $z''(t) = -2z'(t) - 2z(t)$, $z(0) = 2$, $z'(0) = -2$.

Exercise 2.2.15: Let us revisit the Cauchy–Euler equations of [Exercise 2.1.26](#) on page 110. Suppose now that $(b - a)^2 - 4ac < 0$. Find a formula for the general solution of $ax^2y'' + bxy' + cy = 0$. Hint: Note that $x^r = e^{r \ln x}$.

Exercise 2.2.16: Construct an equation such that $y = C_1e^{-2x} \cos(3x) + C_2e^{-2x} \sin(3x)$ is the general solution.

Exercise 2.2.17: Find the solution to the initial value problem

$$y'' + 4y' + 5y = 0 \quad y(0) = 3, \quad y'(0) = -1.$$

Determine a value T so that $|y(x)| < 0.02$ for all $x > T$.

Exercise 2.2.18: Find the solution to the initial value problem

$$y'' + 6y' + 13y = 0 \quad y(0) = 4, \quad y'(0) = 7.$$

Determine a value T so that $|y(x)| < 0.01$ for all $x > T$.

2.3 Repeated Roots and Reduction of Order

Attribution: [JL], §2.1, 2.2.

Learning Objectives

After this section, you will be able to:

- Find the general solution to a second order constant coefficient equation with repeated roots,
- Apply the method of reduction of order to generate a second solution to an equation given one solution, and
- Solve Euler equations using the method of reduction of order.

The last case we have to handle for solving all second order linear constant coefficient equations is the case where $b^2 - 4ac = 0$ in the equation

$$ay'' + by' + cy = 0.$$

When we try to find the characteristic equation and find solutions to this equation, we get a double root at r_1 , so that the characteristic polynomial is $(r - r_1)^2$. For this, we get that $e^{r_1 x}$ is a solution. However, that's the only solution we get. We need to have two linearly independent solutions in order to get the general solution to the differential equation, so we need to find some method to get another solution. The standard method, and the one we apply here is *reduction of order*. Let's see how this works through an example.

Example 2.3.1: Find two linearly independent solutions to the differential equation

$$y'' + 2y' + y = 0.$$

Solution: To start, we find the first solution using our original method. The characteristic equation here is $r^2 + 2r + 1 = 0$, which is $(r + 1)^2$. Therefore, we have a double root at $r = -1$, so that $y_1(x) = e^{-x}$ is a solution.

To find a second solution, the reduction of order method suggests that we try to plug in $y = v(x)e^{-x}$ for an unknown function $v(x)$. The goal is to figure out an equation that v must satisfy to see if this leads us to a second solution to the original equation. We can compute the first two derivatives of $y = v(x)e^{-x}$

$$\begin{aligned} y(x) &= v(x)e^{-x} \\ y'(x) &= v'(x)e^{-x} - v(x)e^{-x} \\ y''(x) &= y''(x)e^{-x} - 2v'(x)e^{-x} + v(x)e^{-x} \end{aligned}$$

and then plug them into the original differential equation

$$\begin{aligned} 0 &= y'' + 2y' + y \\ &= (v''(x)e^{-x} - 2v'(x)e^{-x} + v(x)e^{-x}) + 2(v'(x)e^{-x} - v(x)e^{-x}) + v(x)e^{-x} \\ &= v''(x)e^{-x} + v'(x)(-2e^{-x} + 2e^{-x}) + v(x)(e^{-x} - 2e^{-x} + e^{-x}) \\ &= v''(x)e^{-x} \end{aligned}$$

Since e^{-x} is never zero, this means we must have $v''(x) = 0$. This is still a second order equation, but we know how to solve it. We can integrate both sides twice to get that $v(x) = Ax + B$ for any constants A and B .

Our goal with all of this was to find a solution y of the form $v(x)e^{-x}$. The set up here means that $y = (Ax + B)e^{-x}$ will solve the differential equation. Since we already knew that Be^{-x} was a solution, the new information we gained here was that Axe^{-x} , or in particular, xe^{-x} is a solution to the differential equation. Thus, our two solutions are $y_1(x) = e^{-x}$ and $y_2(x) = xe^{-x}$. \square

Exercise 2.3.1: Check that e^{-x} and xe^{-x} both solve $y'' + 2y' + y = 0$, and that these solutions are linearly independent.

The *reduction of order method* applies more generally to any second order linear homogeneous equation and the goal is the same: use one solution of the differential equation to generate another one. The idea is that if we somehow found y_1 as a solution of $y'' + p(x)y' + q(x)y = 0$ we try a second solution of the form $y_2(x) = y_1(x)v(x)$. We just need to find v . We plug y_2 into the equation:

$$\begin{aligned} 0 &= y_2'' + p(x)y_2' + q(x)y_2 = y_1''v + 2y_1'v' + y_1v'' + p(x)(y_1'v + y_1v') + q(z)y_1v \\ &= y_1v'' + (2y_1' + p(x)y_1)v' + \cancel{(y_1'' + p(x)y_1' + q(x)y_1)}^0 v. \end{aligned}$$

In other words, $y_1v'' + (2y_1' + p(x)y_1)v' = 0$. Using $w = v'$ we have the first order linear equation $y_1w' + (2y_1' + p(x)y_1)w = 0$. After solving this equation for w (integrating factor), we find v by antiderivativing w . We then form y_2 by computing y_1v . For example, suppose we somehow know $y_1 = x$ is a solution to $y'' + x^{-1}y' - x^{-2}y = 0$. The equation for w is then $xw' + 3w = 0$. We find a solution, $w = Cx^{-3}$, and we find an antiderivative $v = \frac{-C}{2x^2}$. Hence $y_2 = y_1v = \frac{-C}{2x}$. Any C works and so $C = -2$ makes $y_2 = 1/x$. Thus, the general solution is $y = C_1x + C_2/x$.

The easiest way to work out these problems is to remember that we need to try $y_2(x) = y_1(x)v(x)$ and find $v(x)$ as we did above. Also, the technique works for higher order equations too: you get to reduce the order for each solution you find.

In summary, for constant coefficient equations with a repeated root, the reduction of order method will always give the equation $v'' = 0$, and so the solution is $v(x) = Ax + B$. Multiplying by the y_1 solution e^{rx} gives that xe^{rx} is the other solution. Therefore, the general solution for repeated root equations is always of the form

$$y = C_1e^{r_1x} + C_2xe^{r_1x}.$$

Example 2.3.2: Find the general solution of

$$y'' - 8y' + 16y = 0.$$

Solution: The characteristic equation is $r^2 - 8r + 16 = (r - 4)^2 = 0$. The equation has a double root $r_1 = r_2 = 4$. The general solution is, therefore,

$$y = (C_1 + C_2x)e^{4x} = C_1e^{4x} + C_2xe^{4x}.$$

Exercise 2.3.2: Check that e^{4x} and xe^{4x} are linearly independent.

That e^{4x} solves the equation is clear. If xe^{4x} solves the equation, then we know we are done. Let us compute $y' = e^{4x} + 4xe^{4x}$ and $y'' = 8e^{4x} + 16xe^{4x}$. Plug in

$$y'' - 8y' + 16y = 8e^{4x} + 16xe^{4x} - 8(e^{4x} + 4xe^{4x}) + 16xe^{4x} = 0.$$

□

In some sense, a doubled root rarely happens. If coefficients are picked randomly, a doubled root is unlikely. There are, however, some natural phenomena where a doubled root does happen, so we cannot just dismiss this case. In addition, there are specific physical applications that involve the double root problem, which we will discuss in Section 2.4. Finally, the solution with a doubled root can be thought of as an approximation of the solution with two roots that are very close together, and the behavior of this solution will approximate “nearby” solutions as well.

2.3.1 Exercises

Exercise 2.3.3: Find the general solution to $y'' + 4y' + 4y = 0$.

Exercise 2.3.4:* Find the general solution to $y'' - 6y' + 9y = 0$.

Exercise 2.3.5: Find the solution to $y'' + 6y' + 9y = 0$ with $y(0) = 3$ and $y'(0) = -1$.

Exercise 2.3.6: Solve $y'' - 8y' + 16y = 0$ for $y(0) = 2$, $y'(0) = 0$.

Exercise 2.3.7: Find the general solution of $y'' = 0$ using the methods of this section.

Exercise 2.3.8: The method of this section applies to equations of other orders than two. We will see higher orders later. Try to solve the first order equation $2y' + 3y = 0$ using the methods of this section.

Exercise 2.3.9 (Euler Equations):* Consider the differential equation $x^2y'' + 3xy' - 3y = 0$.

- a) Verify that $y_1(x) = x$ is a solution.
- b) Use reduction of order to find a second linearly independent solution.
- c) Write out the general solution.

Exercise 2.3.10 (Euler Equations):* Consider the differential equation $x^2y'' + 4xy' - 2y = 0$.

- a) Verify that $y_1(x) = \frac{1}{x}$ is a solution.
- b) Use reduction of order to find a second linearly independent solution.
- c) Write out the general solution.

Exercise 2.3.11 (Euler Equations):* Consider the differential equation $x^2y'' - 6xy' - 10y = 0$.

- a) Verify that $y_1(x) = x^2$ is a solution.
- b) Use reduction of order to find a second linearly independent solution.
- c) Write out the general solution.

Exercise 2.3.12: Find the solution to $y'' - (2\alpha)y' + \alpha^2y = 0$, $y(0) = a$, $y'(0) = b$, where α , a , and b are real numbers.

Exercise 2.3.13 (reduction of order): Suppose y_1 is a solution to $y'' + p(x)y' + q(x)y = 0$. By directly plugging into the equation, show that

$$y_2(x) = y_1(x) \int \frac{e^{-\int p(x) dx}}{(y_1(x))^2} dx$$

is also a solution.

Exercise 2.3.14 (Chebyshev's equation of order 1): Take $(1 - x^2)y'' - xy' + y = 0$.

- a) Show that $y = x$ is a solution.
- b) Use reduction of order to find a second linearly independent solution.
- c) Write down the general solution.

Exercise 2.3.15 (Hermite's equation of order 2): Take $y'' - 2xy' + 4y = 0$.

- a) Show that $y = 1 - 2x^2$ is a solution.
- b) Use reduction of order to find a second linearly independent solution. (It's OK to leave a definite integral in the formula.)
- c) Write down the general solution.

The rest of these exercises can be solved using any of the methods discussed in the last three sections. Pick the appropriate method in order to solve the problem.

Exercise 2.3.16: Find the general solution of $y'' + 5y' - 6y = 0$.

Exercise 2.3.17: Find the general solution of $y'' - 2y' + 2y = 0$.

Exercise 2.3.18: Find the general solution of $y'' + 4y' + 4y = 0$.

Exercise 2.3.19: Find the general solution of $y'' + 4y' + 5y = 0$.

Exercise 2.3.20: Find the solution to $y'' - 6y' + 13y = 0$ with $y(0) = 2$ and $y'(0) = 1$.

Exercise 2.3.21: Find the solution to $y'' + 4y' - 12y = 0$ with $y(0) = -1$ and $y'(0) = 3$.

Exercise 2.3.22: Find the solution to $y'' - 6y' + 9y = 0$ with $y(0) = -4$ and $y'(0) = -1$.

2.4 Mechanical vibrations

Attribution: [JL], §2.4.

Learning Objectives

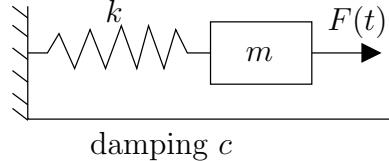
After this section, you will be able to:

- Write second-order differential equations to model physical situations,
- Classify a mechanical oscillation as undamped, underdamped, critically damped, or overdamped, and
- Use the solution to a differential equation to describe the resulting physical motion.

Let us look at some applications of linear second order constant coefficient equations.

2.4.1 Some examples

Our first example is a mass on a spring. Suppose we have a mass $m > 0$ (in kilograms) connected by a spring with spring constant $k > 0$ (in newtons per meter) to a fixed wall. There may be some external force $F(t)$ (in newtons) acting on the mass. Finally, there is some friction measured by $c \geq 0$ (in newton-seconds per meter) as the mass slides along the floor (or perhaps a damper is connected).



Let x be the displacement of the mass ($x = 0$ is the rest position), with x growing to the right (away from the wall). The force exerted by the spring is proportional to the compression of the spring by Hooke's law. Therefore, it is $-kx$ in the negative direction. Similarly the amount of force exerted by friction is proportional to the velocity of the mass. By Newton's second law we know that force equals mass times acceleration and hence $mx'' = F(t) - cx' - kx$ or

$$mx'' + cx' + kx = F(t).$$

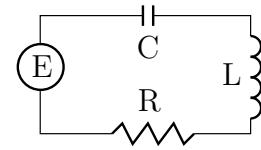
This is a linear second order constant coefficient ODE. We say the motion is

- (i) *forced*, if $F \not\equiv 0$ (if F is not identically zero),
- (ii) *unforced* or *free*, if $F \equiv 0$ (if F is identically zero),
- (iii) *damped*, if $c > 0$, and
- (iv) *undamped*, if $c = 0$.

This system appears in lots of applications even if it does not at first seem like it. Many real-world scenarios can be simplified to a mass on a spring. For example, a bungee jump setup is essentially a mass and spring system (you are the mass). It would be good if someone did the math before you jump off the bridge, right? Let us give two other examples.

Here is an example for electrical engineers. Consider the pictured RLC circuit. There is a resistor with a resistance of R ohms, an inductor with an inductance of L henries, and a capacitor with a capacitance of C farads. There is also an electric source (such as a battery) giving a voltage of $E(t)$ volts at time t (measured in seconds). Let $Q(t)$ be the charge in coulombs on the capacitor and $I(t)$ be the current in the circuit. The relation between the two is $Q' = I$. By elementary principles we find $LI' + RI + Q/C = E$. We differentiate to get

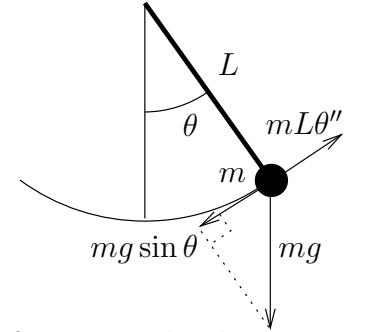
$$LI''(t) + RI'(t) + \frac{1}{C}I(t) = E'(t).$$



This is a nonhomogeneous second order constant coefficient linear equation. As L , R , and C are all positive, this system behaves just like the mass and spring system. Position of the mass is replaced by current. Mass is replaced by inductance, damping is replaced by resistance, and the spring constant is replaced by one over the capacitance. The change in voltage becomes the forcing function—for constant voltage this is an unforced motion.

Our next example behaves like a mass and spring system only approximately. Suppose a mass m hangs on a pendulum of length L . We seek an equation for the angle $\theta(t)$ (in radians). Let g be the force of gravity. Elementary physics mandates that the equation is

$$\theta'' + \frac{g}{L} \sin \theta = 0.$$



Let us derive this equation using Newton's second law: force equals mass times acceleration. The acceleration is $L\theta''$ and mass is m . So $mL\theta''$ has to be equal to the tangential component of the force given by the gravity, which is $mg \sin \theta$ in the opposite direction. So $mL\theta'' = -mg \sin \theta$. The m curiously cancels from the equation.

Now we make our approximation. For small θ we have that approximately $\sin \theta \approx \theta$. This can be seen by looking at the graph. In Figure 2.3 we can see that for approximately $-0.5 < \theta < 0.5$ (in radians) the graphs of $\sin \theta$ and θ are almost the same.

Therefore, when the swings are small, θ is small and we can model the behavior by the simpler linear equation

$$\theta'' + \frac{g}{L} \theta = 0.$$

The errors from this approximation build up. So after a long time, the state of the real-world system might be substantially different from our solution. Also we will see that in a mass-spring system, the amplitude is independent of the period. This is not true for a pendulum. Nevertheless, for reasonably short periods of time and small swings (that is, only small angles θ), the approximation is reasonably good.

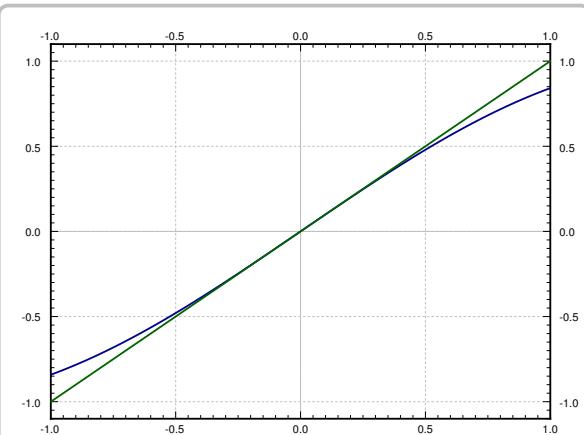


Figure 2.3: The graphs of $\sin \theta$ and θ (in radians).

In real-world problems it is often necessary to make these types of simplifications. We must understand both the mathematics and the physics of the situation to see if the simplification is valid in the context of the questions we are trying to answer.

2.4.2 Free undamped motion

In this section we only consider free or unforced motion, as we do not know yet how to solve nonhomogeneous equations. Let us start with undamped motion where $c = 0$. The equation is

$$mx'' + kx = 0.$$

We divide by m and let $\omega_0 = \sqrt{k/m}$ to rewrite the equation as

$$x'' + \omega_0^2 x = 0.$$

The general solution to this equation is

$$x(t) = A \cos(\omega_0 t) + B \sin(\omega_0 t).$$

By a trigonometric identity

$$A \cos(\omega_0 t) + B \sin(\omega_0 t) = C \cos(\omega_0 t - \gamma),$$

for two different constants C and γ . It is not hard to compute that $C = \sqrt{A^2 + B^2}$ and $\tan \gamma = B/A$. Therefore, we let C and γ be our arbitrary constants and write $x(t) = C \cos(\omega_0 t - \gamma)$.

Exercise 2.4.1: Justify the identity $A \cos(\omega_0 t) + B \sin(\omega_0 t) = C \cos(\omega_0 t - \gamma)$ and verify the equations for C and γ . Hint: Start with $\cos(\alpha - \beta) = \cos(\alpha) \cos(\beta) + \sin(\alpha) \sin(\beta)$ and multiply by C . Then what should α and β be?

While it is generally easier to use the first form with A and B to solve for the initial conditions, the second form is much more natural. The constants C and γ have nice physical interpretation. Write the solution as

$$x(t) = C \cos(\omega_0 t - \gamma).$$

This is a pure-frequency oscillation (a sine wave). The *amplitude* is C , ω_0 is the (angular) *frequency*, and γ is the so-called *phase shift*. The phase shift just shifts the graph left or right. We call ω_0 the *natural (angular) frequency*. This entire setup is called *simple harmonic motion*.

Let us pause to explain the word *angular* before the word *frequency*. The units of ω_0 are radians per unit time, not cycles per unit time as is the usual measure of frequency. Because one cycle is 2π radians, the usual frequency is given by $\frac{\omega_0}{2\pi}$. It is simply a matter of where we put the constant 2π , and that is a matter of taste.

The *period* of the motion is one over the frequency (in cycles per unit time) and hence $\frac{2\pi}{\omega_0}$. That is the amount of time it takes to complete one full cycle.

Example 2.4.1: Suppose that $m = 2\text{ kg}$ and $k = 8\text{ N/m}$. The whole mass and spring setup is sitting on a truck that was traveling at 1 m/s . The truck crashes and hence stops. The mass

was held in place 0.5 meters forward from the rest position. During the crash the mass gets loose. That is, the mass is now moving forward at 1 m/s, while the other end of the spring is held in place. The mass therefore starts oscillating. What is the frequency of the resulting oscillation? What is the amplitude? The units are the mks units (meters-kilograms-seconds).

Solution: The setup means that the mass was at half a meter in the positive direction during the crash and relative to the wall the spring is mounted to, the mass was moving forward (in the positive direction) at 1 m/s. This gives us the initial conditions.

So the equation with initial conditions is

$$2x'' + 8x = 0, \quad x(0) = 0.5, \quad x'(0) = 1.$$

We directly compute $\omega_0 = \sqrt{k/m} = \sqrt{4} = 2$. Hence the angular frequency is 2. The usual frequency in Hertz (cycles per second) is $2/2\pi = 1/\pi \approx 0.318$.

The general solution is

$$x(t) = A \cos(2t) + B \sin(2t).$$

Letting $x(0) = 0.5$ means $A = 0.5$. Then $x'(t) = -2(0.5) \sin(2t) + 2B \cos(2t)$. Letting $x'(0) = 1$ we get $B = 0.5$. Therefore, the amplitude is $C = \sqrt{A^2 + B^2} = \sqrt{0.25 + 0.25} = \sqrt{0.5} \approx 0.707$. The solution is

$$x(t) = 0.5 \cos(2t) + 0.5 \sin(2t).$$

A plot of $x(t)$ is shown in Figure 2.4.

In general, for free undamped motion, a solution of the form

$$x(t) = A \cos(\omega_0 t) + B \sin(\omega_0 t),$$

corresponds to the initial conditions $x(0) = A$ and $x'(0) = \omega_0 B$. Therefore, it is easy to figure out A and B from the initial conditions. The amplitude and the phase shift can then be computed from A and B . In the example, we have already found the amplitude C . Let us compute the phase shift. We know that $\tan \gamma = B/A = 1$. We take the arctangent of 1 and get $\pi/4$ or approximately 0.785. We still need to check if this γ is in the correct quadrant (and add π to γ if it is not). Since both A and B are positive, then γ should be in the first quadrant, $\pi/4$ radians is in the first quadrant, so $\gamma = \pi/4$.

Note: Many calculators and computer software have not only the `atan` function for arctangent, but also what is sometimes called `atan2`. This function takes two arguments, B and A , and returns a γ in the correct quadrant for you.

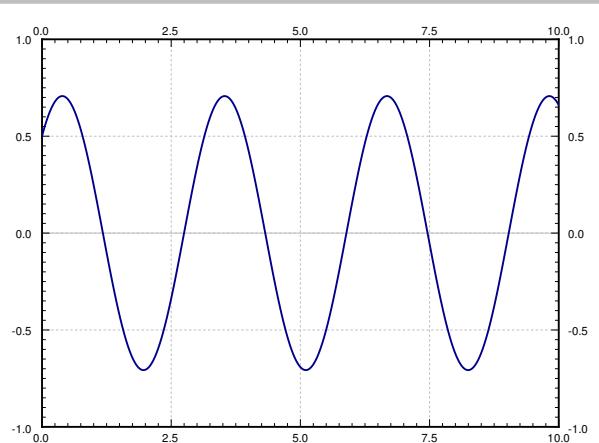


Figure 2.4: Simple undamped oscillation.

2.4.3 Free damped motion

Let us now focus on damped motion. Let us rewrite the equation

$$mx'' + cx' + kx = 0,$$

as

$$x'' + 2px' + \omega_0^2 x = 0,$$

where

$$\omega_0 = \sqrt{\frac{k}{m}}, \quad p = \frac{c}{2m}.$$

The characteristic equation is

$$r^2 + 2pr + \omega_0^2 = 0.$$

Using the quadratic formula we get that the roots are

$$r = -p \pm \sqrt{p^2 - \omega_0^2}.$$

The form of the solution depends on whether we get complex or real roots. We get real roots if and only if the following number is nonnegative:

$$p^2 - \omega_0^2 = \left(\frac{c}{2m}\right)^2 - \frac{k}{m} = \frac{c^2 - 4km}{4m^2}.$$

The sign of $p^2 - \omega_0^2$ is the same as the sign of $c^2 - 4km$. Thus we get real roots if and only if $c^2 - 4km$ is nonnegative, or in other words if $c^2 \geq 4km$.

Overdamping

When $c^2 - 4km > 0$, the system is *overdamped*. In this case, there are two distinct real roots r_1 and r_2 . Both roots are negative: As $\sqrt{p^2 - \omega_0^2}$ is always less than p , then $-p \pm \sqrt{p^2 - \omega_0^2}$ is negative in either case.

The solution is

$$x(t) = C_1 e^{r_1 t} + C_2 e^{r_2 t}.$$

Since r_1, r_2 are negative, $x(t) \rightarrow 0$ as $t \rightarrow \infty$. Thus the mass will tend towards the rest position as time goes to infinity. For a few sample plots for different initial conditions, see Figure 2.5.

No oscillation happens. In fact, the graph crosses the x -axis at most once. To see why,

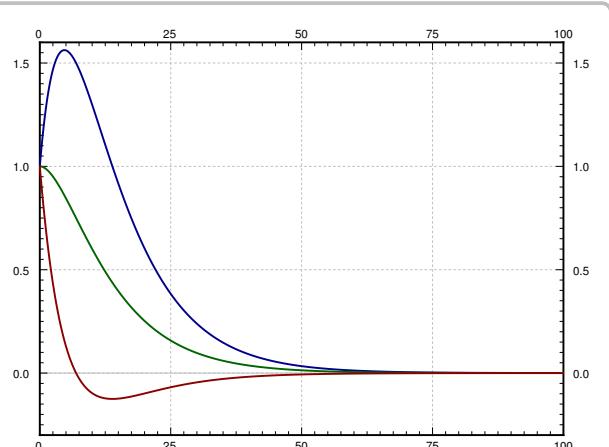


Figure 2.5: Overdamped motion for several different initial conditions.

we try to solve $0 = C_1 e^{r_1 t} + C_2 e^{r_2 t}$. Therefore, $C_1 e^{r_1 t} = -C_2 e^{r_2 t}$ and using laws of exponents we obtain

$$\frac{-C_1}{C_2} = e^{(r_2 - r_1)t}.$$

This equation has at most one solution $t \geq 0$. For some initial conditions the graph never crosses the x -axis, as is evident from the sample graphs.

Example 2.4.2: Suppose the mass is released from rest. That is $x(0) = x_0$ and $x'(0) = 0$. Then

$$x(t) = \frac{x_0}{r_1 - r_2} (r_1 e^{r_2 t} - r_2 e^{r_1 t}).$$

It is not hard to see that this satisfies the initial conditions.

Critical damping

When $c^2 - 4km = 0$, the system is *critically damped*. In this case, there is one root of multiplicity 2 and this root is $-p$. Our solution is

$$x(t) = C_1 e^{-pt} + C_2 t e^{-pt}.$$

The behavior of a critically damped system is very similar to an overdamped system. After all a critically damped system is in some sense a limit of overdamped systems. Even though our models are only approximations of the real world problem, the idea of critical damping can be helpful in optimizing systems. [Figure 2.6](#) shows how the solution to

$$x'' + \gamma x' + x = 0$$

for different values of γ and initial conditions $x(0) = 4$ and $x'(0) = 0$. This solution is critically damped if $\gamma = 2$, as that will give us a repeated root in the characteristic equation. Comparing these solutions, we see that the critically damped solution gets back to equilibrium faster than any of the more overdamped solution. When trying to design a system, if we want it to settle back to the zero point as quickly as possible, then we should try to get as close to critically damped as possible. Even though we are always a little bit underdamped or a little bit overdamped, getting as close as possible will give the best possible result for returning to equilibrium.

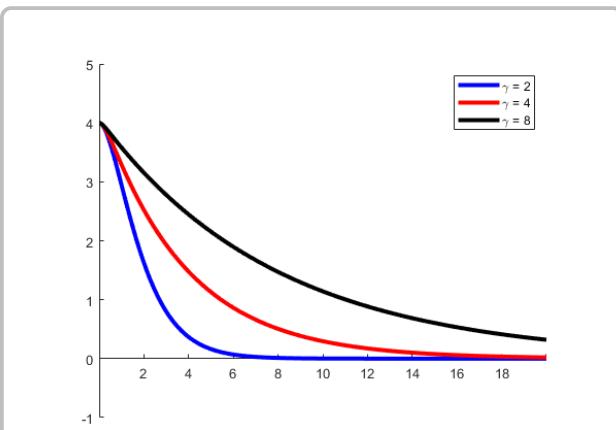


Figure 2.6: Overdamped and critically damped motion for $x'' + \gamma x' + x = 0$ for $\gamma = 2, 4, 8$.

Underdamping

When $c^2 - 4km < 0$, the system is *under-damped*. In this case, the roots are complex.

$$\begin{aligned} r &= -p \pm \sqrt{p^2 - \omega_0^2} \\ &= -p \pm \sqrt{-1} \sqrt{\omega_0^2 - p^2} \\ &= -p \pm i\omega_1, \end{aligned}$$

where $\omega_1 = \sqrt{\omega_0^2 - p^2}$. Our solution is

$$x(t) = e^{-pt}(A \cos(\omega_1 t) + B \sin(\omega_1 t)),$$

or

$$x(t) = Ce^{-pt} \cos(\omega_1 t - \gamma).$$

An example plot is given in Figure 2.7. Note that we still have that $x(t) \rightarrow 0$ as $t \rightarrow \infty$.

The figure also shows the *envelope curves* Ce^{-pt} and $-Ce^{-pt}$. The solution is the oscillating line between the two envelope curves. The envelope curves give the maximum amplitude of the oscillation at any given point in time. For example, if you are bungee jumping, you are really interested in computing the envelope curve as not to hit the concrete with your head.

The phase shift γ shifts the oscillation left or right, but within the envelope curves (the envelope curves do not change if γ changes).

Notice that the angular *pseudo-frequency** becomes smaller when the damping c (and hence p) becomes larger. This makes sense. When we change the damping just a little bit, we do not expect the behavior of the solution to change dramatically. If we keep making c larger, then at some point the solution should start looking like the solution for critical damping or overdamping, where no oscillation happens. So if c^2 approaches $4km$, we want ω_1 to approach 0. Since $\omega_1 = \sqrt{\omega_0^2 - p^2}$ with $p = \frac{c}{2m}$ and $\omega_0 = \sqrt{km}$, we have that

$$\omega_1 = \sqrt{\frac{k}{m} - \frac{c^2}{4m^2}} = \sqrt{\frac{4mk - c^2}{4m^2}},$$

which does go to zero as c^2 gets closer to $4mk$.

On the other hand, when c gets smaller, ω_1 approaches ω_0 (ω_1 is always smaller than ω_0), and the solution looks more and more like the steady periodic motion of the undamped case. The envelope curves become flatter and flatter as c (and hence p) goes to 0.

2.4.4 Exercises

Exercise 2.4.2: Consider a mass and spring system with a mass $m = 2$, spring constant $k = 3$, and damping constant $c = 1$.

- a) Set up and find the general solution of the system.

*We do not call ω_1 a frequency since the solution is not really a periodic function.

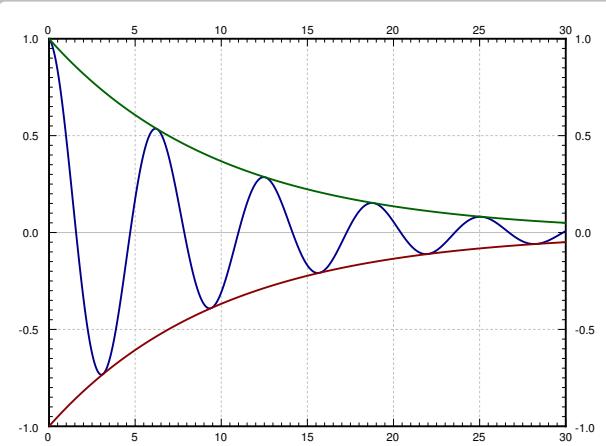


Figure 2.7: Underdamped motion with the envelope curves shown.

- b) Is the system underdamped, overdamped or critically damped?
- c) If the system is not critically damped, find a c that makes the system critically damped.

Exercise 2.4.3: Do [Exercise 2.4.2](#) for $m = 3$, $k = 12$, and $c = 12$.

Exercise 2.4.4: Using the mks units (meters-kilograms-seconds), suppose you have a spring with spring constant 4 N/m . You want to use it to weigh items. Assume no friction. You place the mass on the spring and put it in motion.

- a) You count and find that the frequency is 0.8 Hz (cycles per second). What is the mass?
- b) Find a formula for the mass m given the frequency ω in Hz.

Exercise 2.4.5:* A mass of 2 kilograms is on a spring with spring constant k newtons per meter with no damping. Suppose the system is at rest and at time $t = 0$ the mass is kicked and starts traveling at 2 meters per second. How large does k have to be so that the mass does not go further than 3 meters from the rest position?

Exercise 2.4.6: Suppose we add possible friction to [Exercise 2.4.4](#). Further, suppose you do not know the spring constant, but you have two reference weights 1 kg and 2 kg to calibrate your setup. You put each in motion on your spring and measure the frequency. For the 1 kg weight you measured 1.1 Hz , for the 2 kg weight you measured 0.8 Hz .

- a) Find k (spring constant) and c (damping constant).
- b) Find a formula for the mass in terms of the frequency in Hz. Note that there may be more than one possible mass for a given frequency.
- c) For an unknown object you measured 0.2 Hz , what is the mass of the object? Suppose that you know that the mass of the unknown object is more than a kilogram.

Exercise 2.4.7: Suppose you wish to measure the friction a mass of 0.1 kg experiences as it slides along a floor (you wish to find c). You have a spring with spring constant $k = 5 \text{ N/m}$. You take the spring, you attach it to the mass and fix it to a wall. Then you pull on the spring and let the mass go. You find that the mass oscillates with frequency 1 Hz . What is the friction?

Exercise 2.4.8:* A 5000 kg railcar hits a bumper (a spring) at 1 m/s , and the spring compresses by 0.1 m . Assume no damping.

- a) Find k .
- b) How far does the spring compress when a 10000 kg railcar hits the spring at the same speed?
- c) If the spring would break if it compresses further than 0.3 m , what is the maximum mass of a railcar that can hit it at 1 m/s ?
- d) What is the maximum mass of a railcar that can hit the spring without breaking at 2 m/s ?

Exercise 2.4.9: When attached to a spring, a 2 kg mass stretches the spring by 0.49 m.

- What is the spring constant of this spring? Use 9.8 m/s^2 as the gravity constant.
- This mass is allowed to come to rest, lifted up by 0.4 m and then released. If there is no damping, set up and solve an initial value problem for the position of the mass as a function of time.
- For a next experiment, you attach a dampener of coefficient 16 Ns/m to the system, and give the same initial condition. Set up and solve an initial value problem for the position of the mass. What type of “dampening” would be used to characterize this situation?

Exercise 2.4.10:* A mass of m kg is on a spring with $k = 3 \text{ N/m}$ and $c = 2 \text{ Ns/m}$. Find the mass m_0 for which there is critical damping. If $m < m_0$, does the system oscillate or not, that is, is it underdamped or overdamped?

Exercise 2.4.11:* Suppose we have an RLC circuit with a resistor of 100 milliohms (0.1 ohms), inductor of inductance of 50 millihenries (0.05 henries), and a capacitor of 5 farads, with constant voltage.

- Set up the ODE equation for the current I .
- Find the general solution.
- Solve for $I(0) = 10$ and $I'(0) = 0$.

Exercise 2.4.12: Assume that the system $my'' + cy' + ky = 0$ is either critically or over-damped. Prove that the solution can pass through zero at most once, regardless of initial conditions. Hint: Try to find all values of t for which $y(t) = 0$, given the form of the solution.

2.5 Nonhomogeneous equations

Attribution: [JL], §2.5.

Learning Objectives

After this section, you will be able to:

- Find the corresponding homogeneous equation for a non-homogeneous equation,
- Use the method of undetermined coefficients to solve non-homogeneous equations,
- Use the method of variation of parameters to solve non-homogeneous equations, and
- Solve for the necessary coefficients to solve initial value problems for non-homogeneous equations.

2.5.1 Solving nonhomogeneous equations

We have solved linear constant coefficient homogeneous equations. What about nonhomogeneous linear ODEs? For example, the equations for forced mechanical vibrations, where we add a “forcing” term, which is a function on the right-hand side of the equation. That is, suppose we have an equation such as

$$y'' + 5y' + 6y = 2x + 1. \quad (2.5)$$

We will write $L[y] = 2x + 1$, where $L[y]$ represents the entire left-hand side of $y'' + 5y' + 6y$, when the exact form of the operator is not important. We solve (2.5) in the following manner. First, we find the general solution y_c to the *associated homogeneous equation*

$$y'' + 5y' + 6y = 0. \quad (2.6)$$

We call y_c the *complementary solution*. Next, we find a single *particular solution* y_p to (2.5) in some way. Then

$$y = y_c + y_p$$

is the general solution to (2.5). We have $L[y_c] = 0$ and $L[y_p] = 2x + 1$. As L is a *linear operator* we verify that y is a solution, $L[y] = L[y_c + y_p] = L[y_c] + L[y_p] = 0 + (2x + 1)$. Let us see why we obtain the *general solution*.

Let y_p and \tilde{y}_p be two different particular solutions to (2.5). Write the difference as $w = y_p - \tilde{y}_p$. Then plug w into the left-hand side of the equation to get

$$w'' + 5w' + 6w = (y_p'' + 5y_p' + 6y_p) - (\tilde{y}_p'' + 5\tilde{y}_p' + 6\tilde{y}_p) = (2x + 1) - (2x + 1) = 0.$$

Using the operator notation the calculation becomes simpler. As L is a linear operator we write

$$L[w] = L[y_p - \tilde{y}_p] = L[y_p] - L[\tilde{y}_p] = (2x + 1) - (2x + 1) = 0.$$

So $w = y_p - \tilde{y}_p$ is a solution to (2.6), that is $Lw = 0$. Any two solutions of (2.5) differ by a solution to the homogeneous equation (2.6). The solution $y = y_c + y_p$ includes *all* solutions to (2.5), since y_c is the general solution to the associated homogeneous equation.

Theorem 2.5.1

Let $L[y] = f(x)$ be a linear ODE (not necessarily constant coefficient). Let y_c be the complementary solution (the general solution to the associated homogeneous equation $L[y] = 0$) and let y_p be any particular solution to $L[y] = f(x)$. Then the general solution to $L[y] = f(x)$ is

$$y = y_c + y_p.$$

The moral of the story is that we can find the particular solution in any old way. If we find a different particular solution (by a different method, or simply by guessing), then we still get the same general solution. The formula may look different, and the constants we have to choose to satisfy the initial conditions may be different, but it is the same solution.

2.5.2 Undetermined coefficients

The trick is to somehow, in a smart way, guess one particular solution to (2.5). Note that $2x + 1$ is a polynomial, and the left-hand side of the equation will be a polynomial if we let y be a polynomial of the same degree. Let us try

$$y_p = Ax + B.$$

We plug y_p into the left hand side to obtain

$$\begin{aligned} y_p'' + 5y_p' + 6y_p &= (Ax + B)'' + 5(Ax + B)' + 6(Ax + B) \\ &= 0 + 5A + 6Ax + 6B = 6Ax + (5A + 6B). \end{aligned}$$

So $6Ax + (5A + 6B) = 2x + 1$. Therefore, $A = 1/3$ and $B = -1/9$. That means $y_p = \frac{1}{3}x - \frac{1}{9} = \frac{3x-1}{9}$. Solving the complementary problem (exercise!) we get

$$y_c = C_1 e^{-2x} + C_2 e^{-3x}.$$

Hence the general solution to (2.5) is

$$y = C_1 e^{-2x} + C_2 e^{-3x} + \frac{3x-1}{9}.$$

Now suppose we are further given some initial conditions. For example, $y(0) = 0$ and $y'(0) = 1/3$. First find $y' = -2C_1 e^{-2x} - 3C_2 e^{-3x} + 1/3$. Then

$$0 = y(0) = C_1 + C_2 - \frac{1}{9}, \quad \frac{1}{3} = y'(0) = -2C_1 - 3C_2 + \frac{1}{3}.$$

We solve to get $C_1 = 1/3$ and $C_2 = -2/9$. The particular solution we want is

$$y(x) = \frac{1}{3}e^{-2x} - \frac{2}{9}e^{-3x} + \frac{3x-1}{9} = \frac{3e^{-2x} - 2e^{-3x} + 3x-1}{9}.$$

Exercise 2.5.1: Check that y really solves the equation (2.5) and the given initial conditions.

Note: A common mistake is to solve for constants using the initial conditions with y_c and only add the particular solution y_p after that. That will *not* work. You need to first compute $y = y_c + y_p$ and *only then* solve for the constants using the initial conditions.

A right-hand side consisting of exponentials, sines, and cosines can be handled similarly.

Example 2.5.1: One example of this is

$$y'' + 2y' + 2y = \cos(2x).$$

Solution: Let us find some y_p . We start by guessing the solution includes some multiple of $\cos(2x)$. We may have to also add a multiple of $\sin(2x)$ to our guess since derivatives of cosine are sines. We try

$$y_p = A \cos(2x) + B \sin(2x).$$

We plug y_p into the equation and we get

$$\underbrace{-4A \cos(2x) - 4B \sin(2x)}_{y_p''} + 2 \underbrace{(-2A \sin(2x) + 2B \cos(2x))}_{y_p'} + 2 \underbrace{(A \cos(2x) + 2B \sin(2x))}_{y_p} = \cos(2x),$$

or

$$(-4A + 4B + 2A) \cos(2x) + (-4B - 4A + 2B) \sin(2x) = \cos(2x).$$

The left-hand side must equal to right-hand side. Namely, $-4A + 4B + 2A = 1$ and $-4B - 4A + 2B = 0$. So $-2A + 4B = 1$ and $2A + B = 0$ and hence $A = -1/10$ and $B = 1/5$. So

$$y_p = A \cos(2x) + B \sin(2x) = \frac{-\cos(2x) + 2\sin(2x)}{10}.$$

□

Similarly, if the right-hand side contains exponentials we try exponentials. If

$$L[y] = e^{3x},$$

we try $y = Ae^{3x}$ as our guess and try to solve for A .

When the right-hand side is a multiple of sines, cosines, exponentials, and polynomials, we can use the product rule for differentiation to come up with a guess. We need to guess a form for y_p such that $L[y_p]$ is of the same form, and has all the terms needed to for the right-hand side. For example,

$$L[y] = (1 + 3x^2) e^{-x} \cos(\pi x).$$

For this equation, we guess

$$y_p = (A + Bx + Cx^2) e^{-x} \cos(\pi x) + (D + Ex + Fx^2) e^{-x} \sin(\pi x).$$

We plug in and then hopefully get equations that we can solve for A, B, C, D, E , and F . As you can see this can make for a very long and tedious calculation very quickly. C'est la vie!

There is one hiccup in all this. It could be that our guess actually solves the associated homogeneous equation. That is, suppose we have

$$y'' - 9y = e^{3x}.$$

We would love to guess $y = Ae^{3x}$, but if we plug this into the left-hand side of the equation we get

$$y'' - 9y = 9Ae^{3x} - 9Ae^{3x} = 0 \neq e^{3x}.$$

There is no way we can choose A to make the left-hand side be e^{3x} . The trick in this case is to multiply our guess by x to get rid of duplication with the complementary solution. That is first we compute y_c (solution to $L[y] = 0$)

$$y_c = C_1e^{-3x} + C_2e^{3x},$$

and we note that the e^{3x} term is a duplicate with our desired guess. We modify our guess to $y = Axe^{3x}$ so that there is no duplication anymore. Let us try: $y' = Ae^{3x} + 3Axe^{3x}$ and $y'' = 6Ae^{3x} + 9Axe^{3x}$, so

$$y'' - 9y = 6Ae^{3x} + 9Axe^{3x} - 9Axe^{3x} = 6Ae^{3x}.$$

Thus $6Ae^{3x}$ is supposed to equal e^{3x} . Hence, $6A = 1$ and so $A = 1/6$. We can now write the general solution as

$$y = y_c + y_p = C_1e^{-3x} + C_2e^{3x} + \frac{1}{6}xe^{3x}.$$

It is possible that multiplying by x does not get rid of all duplication. For example,

$$y'' - 6y' + 9y = e^{3x}.$$

The complementary solution is $y_c = C_1e^{3x} + C_2xe^{3x}$. Guessing $y = Axe^{3x}$ would not get us anywhere. In this case we want to guess $y_p = Ax^2e^{3x}$. Basically, we want to multiply our guess by x until all duplication is gone. *But no more!* Multiplying too many times will not work.

Finally, what if the right-hand side has several terms, such as

$$L[y] = e^{2x} + \cos x.$$

In this case we find u that solves $L[u] = e^{2x}$ and v that solves $L[v] = \cos x$ (that is, do each term separately). Then note that if $y = u + v$, then $L[y] = e^{2x} + \cos x$. This is because L is linear; we have $L[y] = L[u + v] = L[u] + L[v] = e^{2x} + \cos x$.

To summarize all of this, we can make a table of the different guesses we should make given the form of the right hand side.

Right hand side	Guess
$a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$	$Ax^n + Bx^{n-1} + \cdots + Nx + P$
e^{ax}	Ae^{ax}
$\cos ax$	$A \cos ax + B \sin ax$
$\sin ax$	$A \cos ax + B \sin ax$

- If there is a product of above terms, guess the product of the guesses. So, for a right hand side of xe^{ax} , the guess should be $(Ax+B)e^{ax}$, and for a right hand side of $x \cos ax$, the guess should be $(Ax+B) \cos ax + (Cx+D) \sin ax$.
- If any part solves the homogeneous problem, multiply that entire component by x until nothing does.

2.5.3 Variation of parameters

The method of undetermined coefficients works for many basic problems that crop up. But it does not work all the time. It only works when the right-hand side of the equation $L[y] = f(x)$ has finitely many linearly independent derivatives, so that we can write a guess that consists of them all. Some equations are a bit tougher. Consider

$$y'' + y = \tan x.$$

Each new derivative of $\tan x$ looks completely different and cannot be written as a linear combination of the previous derivatives. If we start differentiating $\tan x$, we get:

$$\begin{aligned} \sec^2 x, \quad 2 \sec^2 x \tan x, \quad 4 \sec^2 x \tan^2 x + 2 \sec^4 x, \\ 8 \sec^2 x \tan^3 x + 16 \sec^4 x \tan x, \quad 16 \sec^2 x \tan^4 x + 88 \sec^4 x \tan^2 x + 16 \sec^6 x, \quad \dots \end{aligned}$$

This equation calls for a different method. We present the method of *variation of parameters*, which handles any equation of the form $L[y] = f(x)$, provided we can solve certain integrals. For simplicity, we restrict ourselves to second order constant coefficient equations, but the method works for higher order equations just as well (the computations become more tedious). The method also works for equations with nonconstant coefficients, provided we can solve the associated homogeneous equation.

Perhaps it is best to explain this method by example. Let us try to solve the equation

$$L[y] = y'' + y = \tan x.$$

First we find the complementary solution (solution to $L[y_c] = 0$). We get $y_c = C_1 y_1 + C_2 y_2$, where $y_1 = \cos x$ and $y_2 = \sin x$. To find a particular solution to the nonhomogeneous equation we try

$$y_p = y = u_1 y_1 + u_2 y_2,$$

where u_1 and u_2 are *functions* and not constants. We are trying to satisfy $L[y] = \tan x$. That gives us one condition on the functions u_1 and u_2 . Compute (note the product rule!)

$$y' = (u'_1 y_1 + u'_2 y_2) + (u_1 y'_1 + u_2 y'_2).$$

We can still impose one more condition at our discretion to simplify computations (we have two unknown functions, so we should be allowed two conditions). We require that $(u'_1 y_1 + u'_2 y_2) = 0$. This makes computing the second derivative easier.

$$\begin{aligned}y' &= u_1 y'_1 + u_2 y'_2, \\y'' &= (u'_1 y'_1 + u'_2 y'_2) + (u_1 y''_1 + u_2 y''_2).\end{aligned}$$

Since y_1 and y_2 are solutions to $y'' + y = 0$, we find $y''_1 = -y_1$ and $y''_2 = -y_2$. (If the equation was a more general $y'' + p(x)y' + q(x)y = 0$, we would have $y''_i = -p(x)y'_i - q(x)y_i$.) So

$$y'' = (u'_1 y'_1 + u'_2 y'_2) - (u_1 y_1 + u_2 y_2).$$

We have $(u_1 y_1 + u_2 y_2) = y$ and so

$$y'' = (u'_1 y'_1 + u'_2 y'_2) - y,$$

and hence

$$y'' + y = L[y] = u'_1 y'_1 + u'_2 y'_2.$$

For y to satisfy $L[y] = f(x)$ we must have $f(x) = u'_1 y'_1 + u'_2 y'_2$.

What we need to solve are the two equations (conditions) we imposed on u_1 and u_2 :

$$\begin{aligned}u'_1 y_1 + u'_2 y_2 &= 0, \\u'_1 y'_1 + u'_2 y'_2 &= f(x).\end{aligned}$$

We solve for u'_1 and u'_2 in terms of $f(x)$, y_1 and y_2 . We always get these formulas for any $L[y] = f(x)$, where $L[y] = y'' + p(x)y' + q(x)y$. There is a general formula for the solution we could just plug into, but instead of memorizing that, it is better, and easier, to just repeat what we do below. In our case the two equations are

$$\begin{aligned}u'_1 \cos(x) + u'_2 \sin(x) &= 0, \\-u'_1 \sin(x) + u'_2 \cos(x) &= \tan(x).\end{aligned}$$

Hence

$$\begin{aligned}u'_1 \cos(x) \sin(x) + u'_2 \sin^2(x) &= 0, \\-u'_1 \sin(x) \cos(x) + u'_2 \cos^2(x) &= \tan(x) \cos(x) = \sin(x).\end{aligned}$$

And thus

$$\begin{aligned}u'_2 (\sin^2(x) + \cos^2(x)) &= \sin(x), \\u'_2 &= \sin(x), \\u'_1 &= \frac{-\sin^2(x)}{\cos(x)} = -\tan(x) \sin(x).\end{aligned}$$

We integrate u'_1 and u'_2 to get u_1 and u_2 .

$$\begin{aligned} u_1 &= \int u'_1 dx = \int -\tan(x) \sin(x) dx = \frac{1}{2} \ln \left| \frac{\sin(x) - 1}{\sin(x) + 1} \right| + \sin(x), \\ u_2 &= \int u'_2 dx = \int \sin(x) dx = -\cos(x). \end{aligned}$$

So our particular solution is

$$\begin{aligned} y_p &= u_1 y_1 + u_2 y_2 = \frac{1}{2} \cos(x) \ln \left| \frac{\sin(x) - 1}{\sin(x) + 1} \right| + \cos(x) \sin(x) - \cos(x) \sin(x) = \\ &= \frac{1}{2} \cos(x) \ln \left| \frac{\sin(x) - 1}{\sin(x) + 1} \right|. \end{aligned}$$

The general solution to $y'' + y = \tan x$ is, therefore,

$$y = C_1 \cos(x) + C_2 \sin(x) + \frac{1}{2} \cos(x) \ln \left| \frac{\sin(x) - 1}{\sin(x) + 1} \right|.$$

In more generality, we can take the system of equations

$$\begin{aligned} u'_1 y_1 + u'_2 y_2 &= 0, \\ u'_1 y'_1 + u'_2 y'_2 &= f(x). \end{aligned}$$

and solve out for u'_1 and u'_2 using elimination. If we do that, we get that

$$u'_1 = -\frac{y_2(x)f(x)}{y_1(x)y'_2(x) - y'_1(x)y_2(x)} \quad u'_2 = \frac{y_1(x)f(x)}{y_1(x)y'_2(x) - y'_1(x)y_2(x)}.$$

We know that solving the equations this way will work out because we start with the assumption that y_1 and y_2 are linearly independent solutions, and the denominator of both of these fractions is exactly what we know is not zero from this assumption. Therefore, both of these functions can be written this way, we can integrate both of them, and set up our particular solution of the form $y_p(x) = u_1 y_1 + u_2 y_2$ to get

$$y_p(x) = -y_1(x) \int_{x_0}^x \frac{y_2(r)f(r)}{y_1(r)y'_2(r) - y'_1(r)y_2(r)} dr + y_2(x) \int_{x_0}^x \frac{y_1(r)f(r)}{y_1(r)y'_2(r) - y'_1(r)y_2(r)} dr \quad (2.7)$$

where x_0 is any conveniently chosen value (usually zero). Notice the use of r as a dummy variable here to separate the functions being integrated from the actual variable that shows up in the solution. This formula will always work for finding a particular solution to a non-homogeneous equation given that we know the solution to the homogeneous equation, but we may not be able to work out the integrals explicitly. This is the downside of this method, it may always work, but can be very tedious and may not result in nice, closed-form expressions like we might get from other methods.

2.5.4 Exercises

Exercise 2.5.2: Find a particular solution of $y'' - y' - 6y = e^{2x}$.

Exercise 2.5.3: Find a particular solution of $y'' - 4y' + 4y = e^{2x}$.

Exercise 2.5.4:* Find a particular solution to $y'' - y' + y = 2 \sin(3x)$

Exercise 2.5.5: Solve the initial value problem $y'' + 9y = \cos(3x) + \sin(3x)$ for $y(0) = 2$, $y'(0) = 1$.

Exercise 2.5.6: Set up the form of the particular solution but do not solve for the coefficients for $y^{(4)} - 2y''' + y'' = e^x$.

Exercise 2.5.7: Set up the form of the particular solution but do not solve for the coefficients for $y^{(4)} - 2y''' + y'' = e^x + x + \sin x$.

Exercise 2.5.8:* Solve $y'' + 2y' + y = x^2$, $y(0) = 1$, $y'(0) = 2$.

Exercise 2.5.9:

- a) Using variation of parameters find a particular solution of $y'' - 2y' + y = e^x$.
- b) Find a particular solution using undetermined coefficients.
- c) Are the two solutions you found the same? See also [Exercise 2.5.22](#).

Exercise 2.5.10:*

- a) Find a particular solution to $y'' + 2y = e^x + x^3$.
- b) Find the general solution.

Exercise 2.5.11: Find the general solution to $y'' - 3y' - 4y = e^{2t} + 1$.

Exercise 2.5.12: Find the general solution to $y'' - 2y' - 5y = \sin(3t) + 2 \cos(3t)$.

Exercise 2.5.13: Find the general solution to $y'' - 4y' - 21y = e^{-3t} + e^{4t}$.

Exercise 2.5.14: Find the general solution to $y'' - 2y' + y = e^t - t$.

Exercise 2.5.15: Find the general solution to $y'' + 4y = \sec(2t)$ using variation of parameters.

Exercise 2.5.16: Find the solution of the initial value problem $y'' - 2y' - 15y = e^{5t} + 3$, $y(0) = 2$, $y'(0) = -1$.

Exercise 2.5.17: Find the solution of the initial value problem $y'' + 4y' + 5y = \cos(3t) + t$, $y(0) = 0$, $y'(0) = 2$.

Exercise 2.5.18: Find a particular solution of $y'' - 2y' + y = \sin(x^2)$. It is OK to leave the answer as a definite integral.

Exercise 2.5.19: Use variation of parameters to find a particular solution of $y'' - y = \frac{1}{e^x + e^{-x}}$.

Exercise 2.5.20: For an arbitrary constant c find the general solution to $y'' - 2y = \sin(x+c)$.

Exercise 2.5.21: For an arbitrary constant c find a particular solution to $y'' - y = e^{cx}$.

Hint: Make sure to handle every possible real c .

Exercise 2.5.22:

- a) Using variation of parameters find a particular solution of $y'' - y = e^x$.
- b) Find a particular solution using undetermined coefficients.
- c) Are the two solutions you found the same? What is going on?

2.6 Forced oscillations and resonance

Attribution: [JL], §2.6.

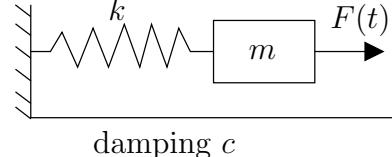
Learning Objectives

After this section, you will be able to:

- Write differential equations to model forced oscillators (like masses on springs),
- Identify when beats, pure resonance, and practical resonance can occur, and
- Use proper terminology around transient and steady periodic solutions when discussing these problems.

Let us return back to the example of a mass on a spring. We examine the case of forced oscillations, which we did not yet handle. That is, we consider the equation

$$mx'' + cx' + kx = F(t),$$



for some nonzero $F(t)$. The setup is again: m is mass, c is friction, k is the spring constant, and $F(t)$ is an external force acting on the mass.

We are interested in periodic forcing, such as noncentered rotating parts, or perhaps loud sounds, or other sources of periodic force.

2.6.1 Undamped forced motion and resonance

First let us consider undamped ($c = 0$) motion. We have the equation

$$mx'' + kx = F_0 \cos(\omega t).$$

This equation has the complementary solution (solution to the associated homogeneous equation)

$$x_c = C_1 \cos(\omega_0 t) + C_2 \sin(\omega_0 t),$$

where $\omega_0 = \sqrt{k/m}$ is the *natural frequency* (angular). It is the frequency at which the system “wants to oscillate” without external interference.

Suppose that $\omega_0 \neq \omega$. We try the solution $x_p = A \cos(\omega t)$ and solve for A . We do not need a sine in our trial solution as after plugging in we only have cosines. If you include a sine, it is fine; you will find that its coefficient is zero (I could not find a second rhyme).

We solve using the method of undetermined coefficients. We find that

$$x_p = \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos(\omega t).$$

We leave it as an exercise to do the algebra required.

The general solution is

$$x = C_1 \cos(\omega_0 t) + C_2 \sin(\omega_0 t) + \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos(\omega t).$$

Written another way

$$x = C \cos(\omega_0 t - \gamma) + \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos(\omega t).$$

The solution is a superposition of two cosine waves at different frequencies.

Example 2.6.1: Take

$$0.5x'' + 8x = 10 \cos(\pi t), \quad x(0) = 0, \quad x'(0) = 0.$$

Solution: Let us compute. First we read off the parameters: $\omega = \pi$, $\omega_0 = \sqrt{8/0.5} = 4$, $F_0 = 10$, $m = 0.5$. The general solution is

$$x = C_1 \cos(4t) + C_2 \sin(4t) + \frac{20}{16 - \pi^2} \cos(\pi t).$$

Solve for C_1 and C_2 using the initial conditions: $C_1 = \frac{-20}{16 - \pi^2}$ and $C_2 = 0$. Hence

$$x = \frac{20}{16 - \pi^2} (\cos(\pi t) - \cos(4t)).$$

Notice the “beating” behavior in Figure 2.8. First use the trigonometric identity

$$2 \sin\left(\frac{A - B}{2}\right) \sin\left(\frac{A + B}{2}\right) = \cos B - \cos A$$

to get

$$x = \frac{20}{16 - \pi^2} \left(2 \sin\left(\frac{4 - \pi}{2}t\right) \sin\left(\frac{4 + \pi}{2}t\right) \right).$$

The function x is a high frequency wave modulated by a low frequency wave. □

The beating behavior can be experienced even more readily by considering a higher frequency and using the resulting function as a sound wave. A sound wave of frequency 400 Hz produces and A4 sound, which is the A above middle C on a piano. This means that the function

$$x_p(t) = \sin(2\pi \cdot 440t)$$

will produce a sound wave equivalent to this A4 sound. In MATLAB, this can be done with the code

```
omega0 = 440*2*pi;
tVals = linspace(0, 5, 5*8192);

testSound = sin(omega0*tVals);
sound(testSound);
```

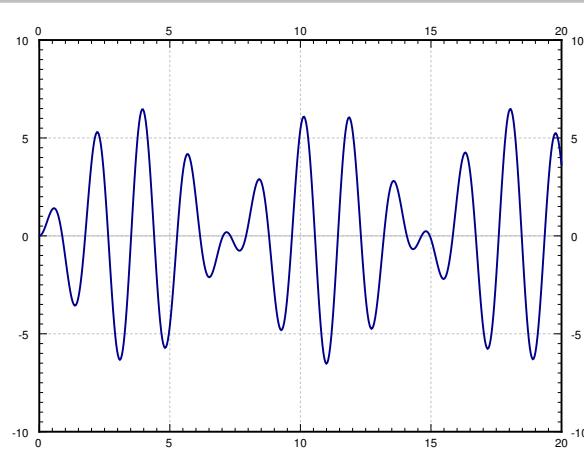


Figure 2.8: Graph of $\frac{20}{16 - \pi^2} (\cos(\pi t) - \cos(4t))$.

which will play this pitch for 5 seconds. Now, we want to see what happens if we take a mass-on-a-spring with this natural frequency and apply a forcing function with frequency close to this value. The following code assumes a forcing function of frequency 444 Hz. The multiple of ω_0 in front of the forcing function is only for scaling purposes; otherwise the resulting sound would be too quiet.

```
omega = 444*2*pi;

syms ys(t);
[V] = odeToVectorField(diff(ys, 2) + omega0^2*ys == omega0*cos(omega*t));
MS = matlabFunction(V, 'vars', {'t', 'Y'});
soln = ode45(MS, [0,10], [0,0]);

ySound = deval(soln, tVals);
ySound = ySound(1, :);
sound(ySound);
```

A graph of the solution $ySound$ can be found in [Figure 2.9](#). This exhibits the beating behavior before on a large scale. The sound played during this code also shows the beating or amplitude modulation that can happen in these sorts of solutions. In terms of tuning instruments, these beats are some of the main things musicians will listen for to know if their instrument is close to the right pitch, but just slightly off.

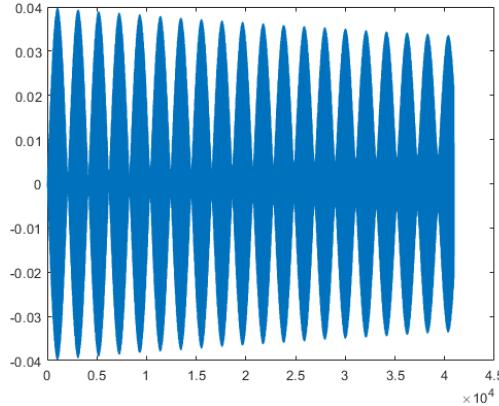


Figure 2.9: Plot of $ySound$ illustrating the beating behavior of interacting sound waves.

Now suppose $\omega_0 = \omega$. Obviously, we cannot try the solution $A \cos(\omega t)$ and then use the method of undetermined coefficients, since we notice that $\cos(\omega t)$ solves the associated homogeneous equation. Therefore, we try $x_p = At \cos(\omega t) + Bt \sin(\omega t)$. This time we need the sine term, since the second derivative of $t \cos(\omega t)$ contains sines. We write the equation

$$x'' + \omega^2 x = \frac{F_0}{m} \cos(\omega t).$$

Plugging x_p into the left-hand side we get

$$2B\omega \cos(\omega t) - 2A\omega \sin(\omega t) = \frac{F_0}{m} \cos(\omega t).$$

Hence $A = 0$ and $B = \frac{F_0}{2m\omega}$. Our particular solution is $\frac{F_0}{2m\omega} t \sin(\omega t)$ and our general solution is

$$x = C_1 \cos(\omega t) + C_2 \sin(\omega t) + \frac{F_0}{2m\omega} t \sin(\omega t).$$

The important term is the last one (the particular solution we found). This term grows without bound as $t \rightarrow \infty$. In fact it oscillates between $\frac{F_0 t}{2m\omega}$ and $-\frac{F_0 t}{2m\omega}$. The first two terms only oscillate between $\pm\sqrt{C_1^2 + C_2^2}$, which becomes smaller and smaller in proportion to the oscillations of the last term as t gets larger. In Figure 2.10 we see the graph with $C_1 = C_2 = 0$, $F_0 = 2$, $m = 1$, $\omega = \pi$.

By forcing the system in just the right frequency we produce very wild oscillations. This kind of behavior is called *resonance* or perhaps *pure resonance*. Sometimes resonance is desired. For example, remember when as a kid you could start swinging by just moving back and forth on the swing seat in the “correct frequency”? You were trying to achieve resonance. The force of each one of your moves was small, but after a while it produced large swings.

On the other hand resonance can be destructive. In an earthquake some buildings collapse while others may be relatively undamaged. This is due to different buildings having different resonance frequencies. So figuring out the resonance frequency can be very important.

A common (but wrong) example of destructive force of resonance is the Tacoma Narrows bridge failure. It turns out there was a different phenomenon at play*.

2.6.2 Damped forced motion and practical resonance

In real life things are not as simple as they were above. There is, of course, some damping. Our equation becomes

$$mx'' + cx' + kx = F_0 \cos(\omega t), \quad (2.8)$$

for some $c > 0$. We solved the homogeneous problem before. We let

$$p = \frac{c}{2m}, \quad \omega_0 = \sqrt{\frac{k}{m}}.$$

*K. Billah and R. Scanlan, *Resonance, Tacoma Narrows Bridge Failure, and Undergraduate Physics Textbooks*, American Journal of Physics, 59(2), 1991, 118–124, <http://www.ketchum.org/billah/Billah-Scanlan.pdf>

We replace equation (2.8) with

$$x'' + 2px' + \omega_0^2 x = \frac{F_0}{m} \cos(\omega t).$$

The roots of the characteristic equation of the associated homogeneous problem are $r_1, r_2 = -p \pm \sqrt{p^2 - \omega_0^2}$. The form of the general solution of the associated homogeneous equation depends on the sign of $p^2 - \omega_0^2$, or equivalently on the sign of $c^2 - 4km$, as before:

$$x_c = \begin{cases} C_1 e^{r_1 t} + C_2 e^{r_2 t} & \text{if } c^2 > 4km, \\ C_1 e^{-pt} + C_2 t e^{-pt} & \text{if } c^2 = 4km, \\ e^{-pt} (C_1 \cos(\omega_1 t) + C_2 \sin(\omega_1 t)) & \text{if } c^2 < 4km, \end{cases}$$

where $\omega_1 = \sqrt{\omega_0^2 - p^2}$. In any case, we see that $x_c(t) \rightarrow 0$ as $t \rightarrow \infty$.

Let us find a particular solution. There can be no conflicts when trying to solve for the undetermined coefficients by trying $x_p = A \cos(\omega t) + B \sin(\omega t)$. Let us plug in and solve for A and B . We get (the tedious details are left to reader)

$$((\omega_0^2 - \omega^2)B - 2\omega p A) \sin(\omega t) + ((\omega_0^2 - \omega^2)A + 2\omega p B) \cos(\omega t) = \frac{F_0}{m} \cos(\omega t).$$

We solve for A and B :

$$\begin{aligned} A &= \frac{(\omega_0^2 - \omega^2)F_0}{m(2\omega p)^2 + m(\omega_0^2 - \omega^2)^2}, \\ B &= \frac{2\omega p F_0}{m(2\omega p)^2 + m(\omega_0^2 - \omega^2)^2}. \end{aligned}$$

We also compute $C = \sqrt{A^2 + B^2}$ to be

$$C = \frac{F_0}{m \sqrt{(2\omega p)^2 + (\omega_0^2 - \omega^2)^2}}.$$

Thus our particular solution is

$$x_p = \frac{(\omega_0^2 - \omega^2)F_0}{m(2\omega p)^2 + m(\omega_0^2 - \omega^2)^2} \cos(\omega t) + \frac{2\omega p F_0}{m(2\omega p)^2 + m(\omega_0^2 - \omega^2)^2} \sin(\omega t).$$

Or in the alternative notation we have amplitude C and phase shift γ where (if $\omega \neq \omega_0$)

$$\tan \gamma = \frac{B}{A} = \frac{2\omega p}{\omega_0^2 - \omega^2}.$$

Hence,

$$x_p = \frac{F_0}{m \sqrt{(2\omega p)^2 + (\omega_0^2 - \omega^2)^2}} \cos(\omega t - \gamma).$$

If $\omega = \omega_0$, then $A = 0$, $B = C = \frac{F_0}{2m\omega p}$, and $\gamma = \pi/2$.

For reasons we will explain in a moment, we call x_c the *transient solution* and denote it by x_{tr} . We call the x_p from above the *steady periodic solution* and denote it by x_{sp} . The general solution is

$$x = x_c + x_p = x_{tr} + x_{sp}.$$

The transient solution $x_c = x_{tr}$ goes to zero as $t \rightarrow \infty$, as all the terms involve an exponential with a negative exponent. So for large t , the effect of x_{tr} is negligible and we see essentially only x_{sp} . Hence the name *transient*. Notice that x_{sp} involves no arbitrary constants, and the initial conditions only affect x_{tr} . Thus, the effect of the initial conditions is negligible after some period of time. We might as well focus on the steady periodic solution and ignore the transient solution. See [Figure 2.11](#) for a graph given several different initial conditions.

The speed at which x_{tr} goes to zero depends on p (and hence c). The bigger p is (the bigger c is), the “faster” x_{tr} becomes negligible. So the smaller the damping, the longer the “transient region.” This is consistent with the observation that when $c = 0$, the initial conditions affect the behavior for all time (i.e. an infinite “transient region”).

Let us describe what we mean by resonance when damping is present. Since there were no conflicts when solving with undetermined coefficient, there is no term that goes to infinity. We look instead at the maximum value of the amplitude of the steady periodic solution. Let C be the amplitude of x_{sp} . If we plot C as a function of ω (with all other parameters fixed), we can find its maximum.

We call the ω that achieves this maximum the *practical resonance frequency*. We call the maximal amplitude $C(\omega)$ the *practical resonance amplitude*. Thus when damping is present we talk of *practical resonance* rather than pure resonance. A sample plot for three different values of c is given in [Figure 2.12](#) on the following page. As you can see the practical resonance amplitude grows as damping gets smaller, and practical resonance can disappear altogether when damping is large.

To find the maximum we need to find the derivative $C'(\omega)$. Computation shows

$$C'(\omega) = \frac{-2\omega(2p^2 + \omega^2 - \omega_0^2)F_0}{m((2\omega p)^2 + (\omega_0^2 - \omega^2)^2)^{3/2}}.$$

This is zero either when $\omega = 0$ or when $2p^2 + \omega^2 - \omega_0^2 = 0$. In other words, $C'(\omega) = 0$ when

$$\omega = \sqrt{\omega_0^2 - 2p^2} \quad \text{or} \quad \omega = 0.$$

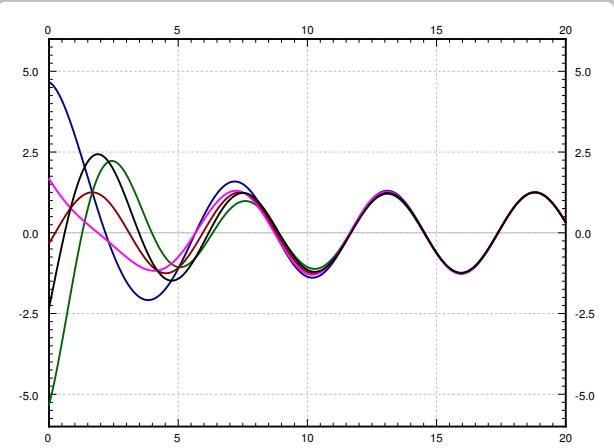


Figure 2.11: Solutions with different initial conditions for parameters $k = 1$, $m = 1$, $F_0 = 1$, $c = 0.7$, and $\omega = 1.1$.

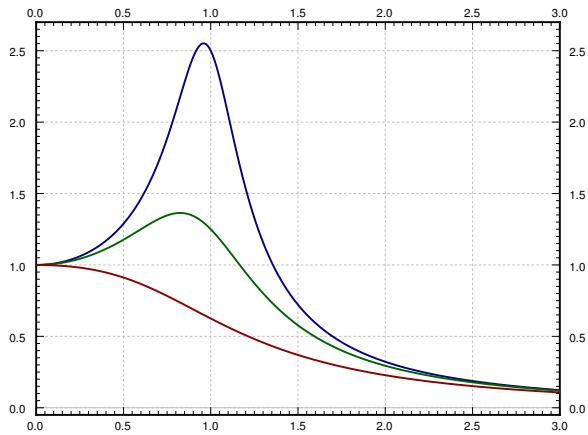


Figure 2.12: Graph of $C(\omega)$ showing practical resonance with parameters $k = 1$, $m = 1$, $F_0 = 1$. The top line is with $c = 0.4$, the middle line with $c = 0.8$, and the bottom line with $c = 1.6$.

If $\omega_0^2 - 2p^2$ is positive, then $\sqrt{\omega_0^2 - 2p^2}$ is the practical resonance frequency (that is the point where $C(\omega)$ is maximal). This follows by the first derivative test for example as then $C'(\omega) > 0$ for small ω in this case. If on the other hand $\omega_0^2 - 2p^2$ is not positive, then $C(\omega)$ achieves its maximum at $\omega = 0$, and there is no practical resonance since we assume $\omega > 0$ in our system. In this case the amplitude gets larger as the forcing frequency gets smaller.

If practical resonance occurs, the frequency is smaller than ω_0 . As the damping c (and hence p) becomes smaller, the practical resonance frequency goes to ω_0 . So when damping is very small, ω_0 is a good estimate of the practical resonance frequency. This behavior agrees with the observation that when $c = 0$, then ω_0 is the resonance frequency.

The main takeaways from this graph here is that the amplitude can be larger than 1, which is the idea of resonance in this case. Based on Hooke's law, we know that a constant force of magnitude F_0 will stretch (or compress) a spring with constant k a length of F_0/k . If we take $F_0 = 1$ and $k = 1$, as is done in Figure 2.12, then the resulting magnitude should be 1. However, if we don't use a constant force of magnitude F_0 , but instead use an oscillatory force with frequency ω of the form $F(t) = F_0 \cos(\omega t)$, we get an amplitude of $C(\omega)$. This graph indicates how the forcing frequency changes the amplitude of the resulting oscillation. Since the amplitude "should" be 1 based on F_0/k , if $C(\omega) > 1$, then the frequency chosen is causing an increase in the amplitude, which is the idea of practical resonance.

Another interesting observation to make is that when $\omega \rightarrow \infty$, then $C \rightarrow 0$. This means that if the forcing frequency gets too high it does not manage to get the mass moving in the mass-spring system. This is quite reasonable intuitively. If we wiggle back and forth really fast while sitting on a swing, we will not get it moving at all, no matter how forceful. Fast vibrations just cancel each other out before the mass has any chance of responding by moving one way or the other.

The behavior is more complicated if the forcing function is not an exact cosine wave, but for example a square wave. A general periodic function will be the sum (superposition) of many cosine waves of different frequencies. The reader is encouraged to come back to this section once we have learned about the Fourier series.

2.6.3 Exercises

Exercise 2.6.1: Write $\cos(3x) - \cos(2x)$ as a product of two sine functions.

Exercise 2.6.2: Write $\cos(5x) - \cos(3x)$ as a product of two sine functions.

Exercise 2.6.3: Write $\cos(3x) - \cos(\pi x)$ as a product of two sine functions.

Exercise 2.6.4: Derive a formula for x_{sp} if the equation is $mx'' + cx' + kx = F_0 \sin(\omega t)$. Assume $c > 0$.

Exercise 2.6.5: Derive a formula for x_{sp} if the equation is $mx'' + cx' + kx = F_0 \cos(\omega t) + F_1 \cos(3\omega t)$. Assume $c > 0$.

Exercise 2.6.6:* Derive a formula for x_{sp} for $mx'' + cx' + kx = F_0 \cos(\omega t) + A$, where A is some constant. Assume $c > 0$.

Exercise 2.6.7: Take $mx'' + cx' + kx = F_0 \cos(\omega t)$. Fix $m > 0$, $k > 0$, and $F_0 > 0$. Consider the function $C(\omega)$. For what values of c (solve in terms of m , k , and F_0) will there be no practical resonance (that is, for what values of c is there no maximum of $C(\omega)$ for $\omega > 0$)?

Exercise 2.6.8: Take $mx'' + cx' + kx = F_0 \cos(\omega t)$. Fix $c > 0$, $k > 0$, and $F_0 > 0$. Consider the function $C(\omega)$. For what values of m (solve in terms of c , k , and F_0) will there be no practical resonance (that is, for what values of m is there no maximum of $C(\omega)$ for $\omega > 0$)?

Exercise 2.6.9:* A mass of 4 kg on a spring with $k = 4 \text{ N/m}$ and a damping constant $c = 1 \text{ Ns/m}$. Suppose that $F_0 = 2 \text{ N}$. Using forcing function $F_0 \cos(\omega t)$, find the ω that causes practical resonance and find the amplitude.

Exercise 2.6.10: An infant is bouncing in a spring chair. The infant has a mass of 8 kg, and the chair functions as a spring with spring constant 72 N/m . The bouncing of the infant applies a force of the form $3 \cos(\omega t)$ for some frequency ω . Assume that the infant starts at rest at the equilibrium position of the chair.

- If there is no dampening coefficient, what frequency would the infant need to force at in order to generate pure resonance?
- Assume that the chair is built with a dampener with coefficient 5 Ns/m . Set up an initial value problem for this situation if the child behaves in the same way.
- Solve this initial value problem.
- There are several options for chairs you can buy. There is the one with dampening coefficient 5 Ns/m , one with 1 Ns/m , and one with 20 Ns/m . Which of these would be most ‘fun’ for the infant? How do you know?

Exercise 2.6.11: A water tower in an earthquake acts as a mass-spring system. Assume that the container on top is full and the water does not move around. The container then acts as the mass and the support acts as the spring, where the induced vibrations are horizontal. The container with water has a mass of $m = 10,000 \text{ kg}$. It takes a force of 1000 newtons to displace the container 1 meter. For simplicity assume no friction. When the earthquake hits the water tower is at rest (it is not moving). The earthquake induces an external force $F(t) = mA\omega^2 \cos(\omega t)$.

- a) What is the natural frequency of the water tower?
- b) If ω is not the natural frequency, find a formula for the maximal amplitude of the resulting oscillations of the water container (the maximal deviation from the rest position). The motion will be a high frequency wave modulated by a low frequency wave, so simply find the constant in front of the sines.
- c) Suppose $A = 1$ and an earthquake with frequency 0.5 cycles per second comes. What is the amplitude of the oscillations? Suppose that if the water tower moves more than 1.5 meter from the rest position, the tower collapses. Will the tower collapse?

Exercise 2.6.12:* Suppose there is no damping in a mass and spring system with $m = 5$, $k = 20$, and $F_0 = 5$. Suppose ω is chosen to be precisely the resonance frequency.

- a) Find ω .
- b) Find the amplitude of the oscillations at time $t = 100$, given the system is at rest at $t = 0$.

2.7 Higher order linear ODEs

Attribution: [JL], §2.3.

Learning Objectives

After this section, you will be able to:

- Find the general solution to a linear, constant coefficient, homogeneous differential equation of higher order and
- Solve non-homogeneous higher order equations using the method of undetermined coefficients.

We briefly study higher order equations. Equations appearing in applications tend to be second order. Higher order equations do appear from time to time, but generally the world around us is “second order.”

The basic results about linear ODEs of higher order are essentially the same as for second order equations, with 2 replaced by n . The important concept of linear independence is somewhat more complicated when more than two functions are involved. For higher order constant coefficient ODEs, the methods developed are also somewhat harder to apply, but we will not dwell on these complications. It is also possible to use the methods for systems of linear equations from [chapter 4](#) to solve higher order constant coefficient equations.

Let us start with a general homogeneous linear equation

$$y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = 0. \quad (2.9)$$

Theorem 2.7.1 (Superposition)

Suppose y_1, y_2, \dots, y_n are solutions of the homogeneous equation (2.9). Then

$$y(x) = C_1y_1(x) + C_2y_2(x) + \cdots + C_ny_n(x)$$

also solves (2.9) for arbitrary constants C_1, C_2, \dots, C_n .

In other words, a *linear combination* of solutions to (2.9) is also a solution to (2.9). We also have the existence and uniqueness theorem for nonhomogeneous linear equations.

Theorem 2.7.2 (Existence and uniqueness)

Suppose p_0 through p_{n-1} , and f are continuous functions on some interval I , a is a number in I , and b_0, b_1, \dots, b_{n-1} are constants. The equation

$$y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = f(x)$$

has exactly one solution $y(x)$ defined on the same interval I satisfying the initial conditions

$$y(a) = b_0, \quad y'(a) = b_1, \quad \dots, \quad y^{(n-1)}(a) = b_{n-1}.$$

2.7.1 Linear independence

When we had two functions y_1 and y_2 we said they were linearly independent if one was not the multiple of the other. Same idea holds for n functions. In this case it is easier to state as follows. The functions y_1, y_2, \dots, y_n are *linearly independent* if the equation

$$c_1y_1 + c_2y_2 + \cdots + c_ny_n = 0$$

has only the trivial solution $c_1 = c_2 = \cdots = c_n = 0$, where the equation must hold for all x . If we can solve equation with some constants where for example $c_1 \neq 0$, then we can solve for y_1 as a linear combination of the others. If the functions are not linearly independent, they are *linearly dependent*.

Example 2.7.1: Show that e^x, e^{2x}, e^{3x} are linearly independent.

Solution: Let us give several ways to show this fact. Many textbooks (including [EP] and [F]) introduce Wronskians, but it is difficult to see why they work and they are not really necessary here.

Let us write down

$$c_1e^x + c_2e^{2x} + c_3e^{3x} = 0.$$

We use rules of exponentials and write $z = e^x$. Hence $z^2 = e^{2x}$ and $z^3 = e^{3x}$. Then we have

$$c_1z + c_2z^2 + c_3z^3 = 0.$$

The left-hand side is a third degree polynomial in z . It is either identically zero, or it has at most 3 zeros. Therefore, it is identically zero, $c_1 = c_2 = c_3 = 0$, and the functions are linearly independent.

Let us try another way. As before we write

$$c_1e^x + c_2e^{2x} + c_3e^{3x} = 0.$$

This equation has to hold for all x . We divide through by e^{3x} to get

$$c_1e^{-2x} + c_2e^{-x} + c_3 = 0.$$

As the equation is true for all x , let $x \rightarrow \infty$. After taking the limit we see that $c_3 = 0$. Hence our equation becomes

$$c_1e^x + c_2e^{2x} = 0.$$

Rinse, repeat!

How about yet another way. We again write

$$c_1e^x + c_2e^{2x} + c_3e^{3x} = 0.$$

We can evaluate the equation and its derivatives at different values of x to obtain equations for c_1, c_2 , and c_3 . Let us first divide by e^x for simplicity.

$$c_1 + c_2e^x + c_3e^{2x} = 0.$$

We set $x = 0$ to get the equation $c_1 + c_2 + c_3 = 0$. Now differentiate both sides

$$c_2 e^x + 2c_3 e^{2x} = 0.$$

We set $x = 0$ to get $c_2 + 2c_3 = 0$. We divide by e^x again and differentiate to get $2c_3 e^x = 0$. It is clear that c_3 is zero. Then c_2 must be zero as $c_2 = -2c_3$, and c_1 must be zero because $c_1 + c_2 + c_3 = 0$.

There is no one best way to do it. All of these methods are perfectly valid. The important thing is to understand why the functions are linearly independent. \square

Example 2.7.2: On the other hand, the functions e^x , e^{-x} , and $\cosh x$ are linearly dependent. Simply apply definition of the hyperbolic cosine:

$$\cosh x = \frac{e^x + e^{-x}}{2} \quad \text{or} \quad 2 \cosh x - e^x - e^{-x} = 0.$$

2.7.2 Constant coefficient higher order ODEs

When we have a higher order constant coefficient homogeneous linear equation, the song and dance is exactly the same as it was for second order. We just need to find more solutions. If the equation is n^{th} order, we need to find n linearly independent solutions. It is best seen by example.

Example 2.7.3: Find the general solution to

$$y''' - 3y'' - y' + 3y = 0. \quad (2.10)$$

Solution: Try: $y = e^{rx}$. We plug in and get

$$\underbrace{r^3 e^{rx}}_{y'''} - 3\underbrace{r^2 e^{rx}}_{y''} - \underbrace{r e^{rx}}_{y'} + 3\underbrace{e^{rx}}_y = 0.$$

We divide through by e^{rx} . Then

$$r^3 - 3r^2 - r + 3 = 0.$$

The trick now is to find the roots. There is a formula for the roots of degree 3 and 4 polynomials but it is very complicated. There is no formula for higher degree polynomials. That does not mean that the roots do not exist. There are always n roots for an n^{th} degree polynomial. They may be repeated and they may be complex. Computers are pretty good at finding roots approximately for reasonable size polynomials.

A good place to start is to plot the polynomial and check where it is zero. We can also simply try plugging in. We just start plugging in numbers $r = -2, -1, 0, 1, 2, \dots$ and see if we get a hit (we can also try complex numbers). Even if we do not get a hit, we may get an indication of where the root is. For example, we plug $r = -2$ into our polynomial and get -15 ; we plug in $r = 0$ and get 3 . That means there is a root between $r = -2$ and $r = 0$, because the sign changed. If we find one root, say r_1 , then we know $(r - r_1)$ is a factor of our polynomial. Polynomial long division can then be used.

A good strategy is to begin with $r = 0, 1$, or -1 . These are easy to compute. Our polynomial has two such roots, $r_1 = -1$ and $r_2 = 1$. There should be 3 roots and the last root is reasonably easy to find. The constant term in a monic* polynomial such as this is the

*The word monic means that the coefficient of the top degree r^d , in our case r^3 , is 1.

multiple of the negations of all the roots because $r^3 - 3r^2 - r + 3 = (r - r_1)(r - r_2)(r - r_3)$. So

$$3 = (-r_1)(-r_2)(-r_3) = (1)(-1)(-r_3) = r_3.$$

You should check that $r_3 = 3$ really is a root. Hence e^{-x} , e^x and e^{3x} are solutions to (2.10). They are linearly independent as can easily be checked, and there are 3 of them, which happens to be exactly the number we need. So the general solution is

$$y = C_1 e^{-x} + C_2 e^x + C_3 e^{3x}.$$

Another possible way to work out this general solution is by factoring the original polynomial. Since we want to solve

$$r^3 - 3r^2 - r + 3 = 0,$$

we can rewrite the polynomial as

$$r^2(r - 3) - 1(r - 3) = 0$$

which factors as

$$(r^2 - 1)(r - 3) = 0.$$

Finally, using difference of two squares on the first factor gives

$$(r - 1)(r + 1)(r - 3) = 0.$$

This gives roots of 1, -1, and 3, and so the same general solution as above.

Suppose we were given some initial conditions $y(0) = 1$, $y'(0) = 2$, and $y''(0) = 3$. Then

$$\begin{aligned} 1 &= y(0) = C_1 + C_2 + C_3, \\ 2 &= y'(0) = -C_1 + C_2 + 3C_3, \\ 3 &= y''(0) = C_1 + C_2 + 9C_3. \end{aligned}$$

It is possible to find the solution by high school algebra, but it would be a pain. The sensible way to solve a system of equations such as this is to use matrix algebra, see § 4.2 or Chapter 3. For now we note that the solution is $C_1 = -1/4$, $C_2 = 1$, and $C_3 = 1/4$. The specific solution to the ODE is

$$y = \frac{-1}{4} e^{-x} + e^x + \frac{1}{4} e^{3x}.$$

]

Next, suppose that we have real roots, but they are repeated. Let us say we have a root r repeated k times. In the spirit of the second order solution, and for the same reasons, we have the solutions

$$e^{rx}, \quad xe^{rx}, \quad x^2 e^{rx}, \quad \dots, \quad x^{k-1} e^{rx}.$$

We take a linear combination of these solutions to find the general solution.

Example 2.7.4: Solve

$$y^{(4)} - 3y''' + 3y'' - y' = 0.$$

Solution: We note that the characteristic equation is

$$r^4 - 3r^3 + 3r^2 - r = 0.$$

By inspection we note that $r^4 - 3r^3 + 3r^2 - r = r(r-1)^3$. Hence the roots given with multiplicity are $r = 0, 1, 1, 1$. Thus the general solution is

$$y = \underbrace{(C_1 + C_2x + C_3x^2)e^x}_{\text{terms coming from } r=1} + \underbrace{C_4}_{\text{from } r=0}.$$

]

Example 2.7.5: Find the general solution of

$$y''' + 2y'' - 5y' - 6y = 0$$

Solution: The characteristic equation for this example is

$$r^3 + 2r^2 - 5r - 6 = 0.$$

There is no convenient factoring by grouping or other quick formula to get to the roots here. The best hope we have is to try to guess the roots and see if we come up with anything. Once we get one root, we'll be able to factor a term out and get down to a quadratic equation, where the quadratic formula will give us the other two roots.

The properties of polynomials tell us that all rational roots of this polynomial must be factors of $\frac{-6}{1}$ or -6 . Thus, the options are ± 1 , ± 2 , and ± 3 . At this point, the best bet is to start guessing and see if we can find one. Let's start with 1. Plugging this into the polynomial gives

$$1^3 + 2(1)^2 - 5(1) - 6 = -8 \neq 0.$$

Trying -1 next, we get

$$(-1)^3 + 2(-1)^2 - 5(-1) - 6 = -1 + 2 + 5 - 6 = 0.$$

Therefore, -1 is a root, and so $(r+1)$ is a factor of this polynomial.

We can then use synthetic (or long) division to see that

$$r^3 + 2r^2 - 5r - 6 = (r+1)(r^2 + r - 6).$$

For the quadratic, we can either use the quadratic formula, or just recognize that this factors as $(r-2)(r+3)$ to get that the characteristic equation factors as

$$(r+1)(r-2)(r+3) = 0.$$

Therefore, the roots are $-1, 2$ and -3 , so that the general solution to the differential equation is

$$y(x) = C_1e^{-x} + C_2e^{2x} + C_3e^{-3x}.$$

]

For more information on synthetic division and finding roots of polynomials, see Appendix B.1.

The case of complex roots is similar to second order equations. Complex roots always come in pairs $r = \alpha \pm i\beta$. Suppose we have two such complex roots, each repeated k times. The corresponding solution is

$$(C_0 + C_1x + \cdots + C_{k-1}x^{k-1}) e^{\alpha x} \cos(\beta x) + (D_0 + D_1x + \cdots + D_{k-1}x^{k-1}) e^{\alpha x} \sin(\beta x).$$

where $C_0, \dots, C_{k-1}, D_0, \dots, D_{k-1}$ are arbitrary constants.

Example 2.7.6: Solve

$$y^{(4)} - 4y''' + 8y'' - 8y' + 4y = 0.$$

Solution: The characteristic equation is

$$\begin{aligned} r^4 - 4r^3 + 8r^2 - 8r + 4 &= 0, \\ (r^2 - 2r + 2)^2 &= 0, \\ ((r - 1)^2 + 1)^2 &= 0. \end{aligned}$$

Hence the roots are $1 \pm i$, both with multiplicity 2. Hence the general solution to the ODE is

$$y = (C_1 + C_2x) e^x \cos x + (C_3 + C_4x) e^x \sin x.$$

The way we solved the characteristic equation above is really by guessing or by inspection. It is not so easy in general. We could also have asked a computer or an advanced calculator for the roots. □

2.7.3 Non-Homogeneous Equations

Just like for second order equation, we can solve higher order non-homogeneous equations. The theory is the same; if we can find any single solution to the non-homogeneous problem, then the general solution of the non-homogeneous problem is this single solution plus the general solution to the corresponding homogeneous problem. The trick comes down to finding this single solution, and undetermined coefficients is the main method here.

In using undetermined coefficients, the guesses we want to make are the same as for second order equations. The only way it really gets more complicated is that now it is possible for any exponential or trigonometric function to be a solution to the homogeneous problem, and so more things will need to be multiplied by x in order to get the appropriate guess for the non-homogeneous solution.

Example 2.7.7: Find the general solution to

$$y''' + 2y'' - 5y' - 6y = 3e^{2x} + e^{4x}.$$

Solution: We found the general solution of the homogeneous problem in [Example 2.7.5](#), which is

$$y(x) = C_1e^{-x} + C_2e^{2x} + C_3e^{-3x}.$$

Now, to solve the non-homogeneous problem, we use the method of undetermined coefficients. Since the non-homogeneous part of the equation has terms of the form e^{2x} and e^{4x} , we would want to guess

$$y_p(x) = Ae^{2x} + Be^{4x}.$$

However, e^{2x} solves the homogeneous problem, so we need to multiply it by x , making our actual guess become

$$y_p(x) = Axe^{2x} + Be^{4x}.$$

In order to plug this in, we need to take three derivatives of this guess, which are

$$\begin{aligned} y_p(x) &= Axe^{2x} + Be^{4x} \\ y'_p(x) &= Ae^{2x} + 2Axe^{2x} + 4Be^{4x} \\ y''_p(x) &= 4Ae^{2x} + 4Axe^{2x} + 16Be^{4x} \\ y'''_p(x) &= 12Ae^{2x} + 8Axe^{2x} + 64Be^{4x} \end{aligned}$$

By putting this into the non-homogeneous equation we want to solve, we get

$$\begin{aligned} (12Ae^{2x} + 8Axe^{2x} + 64Be^{4x}) + 2(4Ae^{2x} + 4Axe^{2x} + 16Be^{4x}) \\ - 5(Ae^{2x} + 2Axe^{2x} + 4Be^{4x}) - 6(Axe^{2x} + Be^{4x}) = 3e^{2x} + e^{4x}. \end{aligned}$$

Simplifying the left hand side of this expression gives

$$15Ae^{2x} + 70Be^{4x} = 3e^{2x} + e^{4x}.$$

To satisfy this equation, we want to set $A = \frac{1}{5}$ and $B = \frac{1}{70}$. Therefore, the general solution to the non-homogeneous problem is

$$y(x) = C_1e^{-x} + C_2e^{2x} + C_3e^{-3x} + \frac{1}{5}xe^{2x} + \frac{1}{70}e^{4x}.$$

□

Example 2.7.8: Determine the form of the guess using undetermined coefficients for finding a particular solution of the non-homogeneous problem

$$y^{(9)} + y^{(8)} - 2y^{(5)} - 2y^{(4)} + y' + y = e^x + 3e^{-x} + \sin(x) + 2x.$$

Solution: To determine the guess, we need to first find the solution to the homogeneous equations. The characteristic equation of the homogeneous equation is

$$r^9 + r^8 - 2r^5 - 2r^4 + r + 1 = 0.$$

We could use the root guessing method for this example, and all rational roots must be ± 1 . However, that method is not great for polynomials that are of degree higher than around 3 or 4. So, we'll want to use some other technique to find all of the root.

If we start by grouping pairs of terms, we can rewrite this polynomial as

$$r^8(r+1) - 2r^4(r+1) + 1(r+1) = 0$$

so that it can be rewritten as

$$(r + 1)(r^8 - 2r^4 + 1) = 0.$$

The second factor looks a lot like

$$(s - 1)^2 = s^2 - 2s + 1$$

if we take $s = r^4$. Since

$$(r^4 - 1) = (r^2 + 1)(r^2 - 1) = (r^2 + 1)(r + 1)(r - 1)$$

using difference of squares twice. Thus, the entire characteristic equation can be written as

$$(r + 1)(r^4 - 1)^2 = (r + 1)[(r^2 + 1)(r + 1)(r - 1)]^2 = (r + 1)^3(r - 1)^2(r^2 + 1)^2.$$

Therefore, we have a triple root at -1 , a double root at 1 , and two copies of $(r^2 + 1)$, which has a root of i , corresponding to solutions $\sin(x)$ and $\cos(x)$. Putting all of this together, the general solution to the homogeneous equation is

$$y_c(x) = (C_1 + C_2x + C_3x^2)e^{-x} + (C_4 + C_5x)e^x + (C_6 + C_7x)\sin(x) + (C_8 + C_9x)\cos(x).$$

This has 9 unknown constants in it, which is expected from the ninth order equation.

Now, we need to figure out the appropriate guess for the non-homogeneous solution. Since the non-homogeneous part of the equation is $e^x + 3e^{-x} + \sin x + 2x$, the base guess would be of the form

$$Ae^x + Be^{-x} + C\sin x + D\cos x + Ex + F$$

because we always need to include both $\sin(x)$ and $\cos(x)$ whenever either of them appear. However, we need to factor in what terms show up in the homogeneous solution. For instance, the e^x term has a term with 1 and x in the homogeneous solution, we need to include the next one up in our guess for the solution to the non-homogeneous problem. Taking this into account for all terms gives the desired guess as

$$y_p(x) = Ax^2e^x + Bx^3e^{-x} + Cx^2\sin(x) + Dx^2\cos(x) + Ex + F.$$

—

There is also an extension of variation of parameters to higher order equations. However, the fact that there are more terms in the solution means that the form of the expression is much more complicated than for second order, and is not worth looking into or trying to remember. The easier way to handle these situations using variation of parameters is by converting the higher order equation into a first order system and applying the methods there, which will be covered in § 4.1 and § 4.6 respectively.

2.7.4 Exercises

Exercise 2.7.1: Find the general solution for $y''' - y'' + y' - y = 0$.

Exercise 2.7.2:* Find the general solution of $y^{(5)} - y^{(4)} = 0$.

Exercise 2.7.3: Find the general solution for $y^{(4)} - 5y''' + 6y'' = 0$.

Exercise 2.7.4: Find the general solution for $y''' + 2y'' + 2y' = 0$.

Exercise 2.7.5: Suppose the characteristic equation for an ODE is $(r - 1)^2(r - 2)^2 = 0$.

- a) Find such a differential equation.
- b) Find its general solution.

Exercise 2.7.6: Suppose that a fourth order equation has a solution $y = 2e^{4x}x \cos x$.

- a) Find such an equation.
- b) Find the initial conditions that the given solution satisfies.

Exercise 2.7.7:* Suppose that the characteristic equation of a third order differential equation has roots $\pm 2i$ and 3.

- a) What is the characteristic equation?
- b) Find the corresponding differential equation.
- c) Find the general solution.

Exercise 2.7.8: Find the general solution for the equation of [Exercise 2.7.6](#).

Exercise 2.7.9:* Find the general solution of

$$y^{(4)} - y''' - 5y'' - 23y' - 20y = 0.$$

Exercise 2.7.10: Find the general solution of

$$y''' - 6y'' + 13y' - 10y = 4e^x + 5e^{3x} - 20.$$

Exercise 2.7.11: Find the general solution of

$$y''' - 3y' + 2y = 2e^x - e^{3x}.$$

Exercise 2.7.12: Find the general solution of

$$y''' + 2y'' + y' + 2y = 3\cos(x) + x.$$

Exercise 2.7.13: Find the general solution of

$$y^{(4)} + 2y'' + y = 4x\cos(x) - e^{3x} + 1$$

Hint: Remember, the guess needs to make sure that no terms in it solve the homogeneous equation.

Exercise 2.7.14: Let $f(x) = e^x - \cos x$, $g(x) = e^x + \cos x$, and $h(x) = \cos x$. Are $f(x)$, $g(x)$, and $h(x)$ linearly independent? If so, show it, if not, find a linear combination that works.

Exercise 2.7.15: Let $f(x) = 0$, $g(x) = \cos x$, and $h(x) = \sin x$. Are $f(x)$, $g(x)$, and $h(x)$ linearly independent? If so, show it, if not, find a linear combination that works.

Exercise 2.7.16:* Are e^x , e^{x+1} , e^{2x} , $\sin(x)$ linearly independent? If so, show it, if not find a linear combination that works.

Exercise 2.7.17: Are x , x^2 , and x^4 linearly independent? If so, show it, if not, find a linear combination that works.

Exercise 2.7.18: Are e^x , xe^x , and x^2e^x linearly independent? If so, show it, if not, find a linear combination that works.

Exercise 2.7.19:* Are $\sin(x)$, x , $x\sin(x)$ linearly independent? If so, show it, if not find a linear combination that works.

Exercise 2.7.20:* Solve $1001y''' + 3.2y'' + \pi y' - \sqrt{4}y = 0$, $y(0) = 0$, $y'(0) = 0$, $y''(0) = 0$.

Exercise 2.7.21: Find an equation such that $y = xe^{-2x} \sin(3x)$ is a solution.

Exercise 2.7.22:* Find an equation such that $y = \cos(x)$, $y = \sin(x)$, $y = e^x$ are solutions.

Chapter 3

Linear algebra

3.1 Vectors, mappings, and matrices

Attribution: [JL], §A.1.

Learning Objectives

After this section, you will be able to:

- Express n-tuples of numbers as vectors,
- Perform operations on vectors, and
- Understand how linear maps on vectors give rise to matrices.

In real life, there is most often more than one variable. We wish to organize dealing with multiple variables in a consistent manner, and in particular organize dealing with linear equations and linear mappings, as those both rather useful and rather easy to handle. Mathematicians joke that “to an engineer every problem is linear, and everything is a matrix.” And well, they (the engineers) are not wrong. Quite often, solving an engineering problem is figuring out the right finite-dimensional linear problem to solve, which is then solved with some matrix manipulation. Most importantly, linear problems are the ones that we know how to solve, and we have many tools to solve them. For engineers, mathematicians, physicists, and anybody in a technical field it is absolutely vital to learn linear algebra.

As motivation, suppose we wish to solve

$$\begin{aligned}x - y &= 2, \\ 2x + y &= 4,\end{aligned}$$

for x and y , that is, find numbers x and y such that the two equations are satisfied. Let us perhaps start by adding the equations together to find

$$x + 2x - y + y = 2 + 4, \quad \text{or} \quad 3x = 6.$$

In other words, $x = 2$. Once we have that, we plug in $x = 2$ into the first equation to find $2 - y = 2$, so $y = 0$. OK, that was easy. What is all this fuss about linear equations. Well,

try doing this if you have 5000 unknowns*. Also, we may have such equations not of just numbers, but of functions and derivatives of functions in differential equations. Clearly we need a more systematic way of doing things. A nice consequence of making things systematic and simpler to write down is that it becomes easier to have computers do the work for us. Computers are rather stupid, they do not think, but are very good at doing lots of repetitive tasks precisely, as long as we figure out a systematic way for them to perform the tasks.

3.1.1 Vectors and operations on vectors

Consider n real numbers as an n -tuple:

$$(x_1, x_2, \dots, x_n).$$

The set of such n -tuples is the so-called *n -dimensional space*, often denoted by \mathbb{R}^n . Sometimes we call this the *n -dimensional euclidean space*[†]. In two dimensions, \mathbb{R}^2 is called the *cartesian plane*[‡], and in three dimensions, it is the same “3-dimensional space” that is dealt with in multivariable calculus. Each such n -tuple represents a point in the n -dimensional space. For example, the point $(1, 2)$ in the plane \mathbb{R}^2 is one unit to the right and two units up from the origin.

When we do algebra with these n -tuples of numbers we call them *vectors*[§]. Mathematicians are keen on separating what is a vector and what is a point of the space or in the plane, and it turns out to be an important distinction, however, for the purposes of linear algebra we can think of everything being represented by a vector. A way to think of a vector, which is especially useful in calculus and differential equations, is an arrow. It is an object that has a *direction* and a *magnitude*. For example, the vector $(1, 2)$ is the arrow from the origin to the point $(1, 2)$ in the plane. The magnitude is the length of the arrow. See [Figure 3.1](#) on the facing page. If we think of vectors as arrows, the arrow doesn’t always have to start at the origin. If we do move it around, however, it should always keep the same direction and the same magnitude.

As vectors are arrows, when we want to give a name to a vector, we draw a little arrow above it:

$$\vec{x}$$

Another popular notation is \mathbf{x} , although we will use the little arrows. It may be easy to write a bold letter in a book, but it is not so easy to write it by hand on paper or on the board. Mathematicians often don’t even write the arrows. A mathematician would write x and just remember that x is a vector and not a number. Just like you remember that Bob is your uncle, and you don’t have to keep repeating “Uncle Bob” and you can just say “Bob.” In this book, however, we will call Bob “Uncle Bob” and write vectors with the little arrows.

*One of the downsides of making everything look like a linear problem is that the number of variables tends to become huge.

[†]Named after the ancient Greek mathematician [Euclid of Alexandria](#) (around 300 BC), possibly the most famous of mathematicians; even small towns often have Euclid Street or Euclid Avenue.

[‡]Named after the French mathematician [René Descartes](#) (1596–1650). It is “cartesian” as his name in Latin is Renatus Cartesius.

[§]A common notation to distinguish vectors from points is to write $(1, 2)$ for the point and $\langle 1, 2 \rangle$ for the vector. We write both as $(1, 2)$.

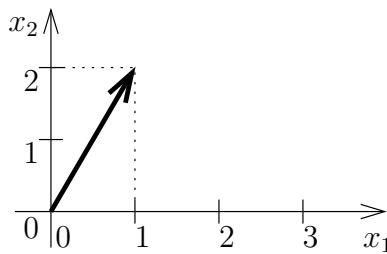


Figure 3.1: The vector $(1, 2)$ drawn as an arrow from the origin to the point $(1, 2)$.

The *magnitude* can be computed using Pythagorean theorem. The vector $(1, 2)$ drawn in the figure has magnitude $\sqrt{1^2 + 2^2} = \sqrt{5}$. The magnitude is denoted by $\|\vec{x}\|$, and, in any number of dimensions, it can be computed in the same way:

$$\|\vec{x}\| = \|(x_1, x_2, \dots, x_n)\| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}.$$

For reasons that will become clear in the next section, we often write vectors as so-called *column vectors*:

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Don't worry. It is just a different way of writing the same thing, and it will be useful later. For example, the vector $(1, 2)$ can be written as

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

The fact that we write arrows above vectors allows us to write several vectors \vec{x}_1 , \vec{x}_2 , etc., without confusing these with the components of some other vector \vec{x} .

So where is the *algebra* from *linear algebra*? Well, arrows can be added, subtracted, and multiplied by numbers. First we consider *addition*. If we have two arrows, we simply move along one, and then along the other. See Figure 3.2.

It is rather easy to see what it does to the numbers that represent the vectors. Suppose we want to add $(1, 2)$ to $(2, -3)$ as in the figure. So we travel along $(1, 2)$ and then we travel along $(2, -3)$. What we did was travel

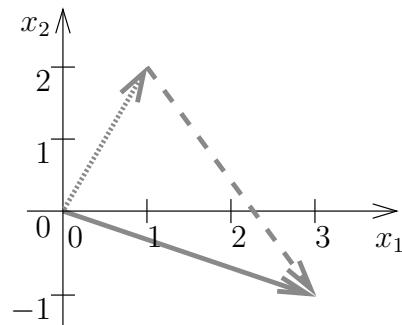


Figure 3.2: Adding the vectors $(1, 2)$, drawn dotted, and $(2, -3)$, drawn dashed. The result, $(3, -1)$, is drawn as a solid arrow.

one unit right, two units up, and then we travelled two units right, and three units down (the negative three). That means that we ended up at $(1 + 2, 2 + (-3)) = (3, -1)$. And that's how addition always works:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}.$$

Subtracting is similar. What $\vec{x} - \vec{y}$ means visually is that we first travel along \vec{x} , and then we travel backwards along \vec{y} . See [Figure 3.3](#). It is like adding $\vec{x} + (-\vec{y})$ where $-\vec{y}$ is the arrow we obtain by erasing the arrow head from one side and drawing it on the other side, that is, we reverse the direction. In terms of the numbers, we simply go backwards in both directions, so we negate both numbers. For example, if \vec{y} is $(-2, 1)$, then $-\vec{y}$ is $(2, -1)$.

Another intuitive thing to do to a vector is to *scale* it. We represent this by multiplication of a number with a vector. Because of this, when we wish to distinguish between vectors and numbers, we call the numbers *scalars*. For example, suppose we want to travel three times further. If the vector is $(1, 2)$, travelling 3 times further means going 3 units to the right and 6 units up, so we get the vector $(3, 6)$. We just multiply each number in the vector by 3. If α is a number, then

$$\alpha \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix}.$$

Scaling (by a positive number) multiplies the magnitude and leaves direction untouched. The magnitude of $(1, 2)$ is $\sqrt{5}$. The magnitude of 3 times $(1, 2)$, that is, $(3, 6)$, is $3\sqrt{5}$.

When the scalar is negative, then when we multiply a vector by it, the vector is not only scaled, but it also switches direction. So multiplying $(1, 2)$ by -3 means we should go 3 times further but in the opposite direction, so 3 units to the left and 6 units down, or in other words, $(-3, -6)$. As we mentioned above, $-\vec{y}$ is a reverse of \vec{y} , and this is the same as $(-1)\vec{y}$.

In [Figure 3.4](#) on the next page, you can see a couple of examples of what scaling a vector means visually.

We put all of these operations together to work out more complicated expressions. Let us compute a small example:

$$3 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 2 \begin{bmatrix} -4 \\ -1 \end{bmatrix} - 3 \begin{bmatrix} -2 \\ 2 \end{bmatrix} = \begin{bmatrix} 3(1) + 2(-4) - 3(-2) \\ 3(2) + 2(-1) - 3(2) \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

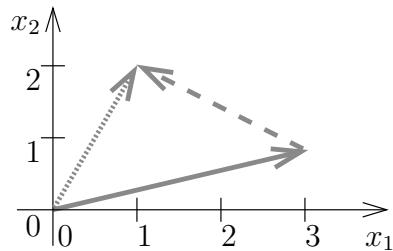


Figure 3.3: Subtraction, the vector $(1, 2)$, drawn dotted, minus $(-2, 1)$, drawn dashed. The result, $(3, 1)$, is drawn as a solid arrow.

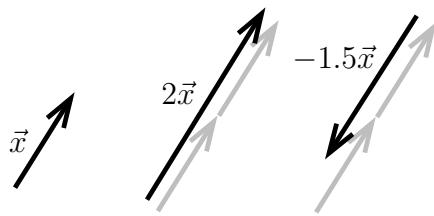


Figure 3.4: A vector \vec{x} , the vector $2\vec{x}$ (same direction, double the magnitude), and the vector $-1.5\vec{x}$ (opposite direction, 1.5 times the magnitude).

As we said a vector is a direction and a magnitude. Magnitude is easy to represent, it is just a number. The *direction* is usually given by a vector with magnitude one. We call such a vector a *unit vector*. That is, \vec{u} is a unit vector when $\|\vec{u}\| = 1$. For example, the vectors $(1, 0)$, $(1/\sqrt{2}, 1/\sqrt{2})$, and $(0, -1)$ are all unit vectors.

To represent the direction of a vector \vec{x} , we need to find the unit vector in the same direction. To do so, we simply rescale \vec{x} by the reciprocal of the magnitude, that is $\frac{1}{\|\vec{x}\|}\vec{x}$, or more concisely $\frac{\vec{x}}{\|\vec{x}\|}$.

For example, the unit vector in the direction of $(1, 2)$ is the vector

$$\frac{1}{\sqrt{1^2 + 2^2}}(1, 2) = \left(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right).$$

3.1.2 Matrices

The next object we need to define here is a *matrix*.

Definition 3.1.1

In general, an $m \times n$ matrix A is a rectangular array of mn numbers,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

An $m \times n$ matrix indicates that it will have m rows and n columns.

Matrices, just like vectors, are generally written with square brackets on the outside, although some books will use parentheses for this. The convention for notation is that matrices will be denoted by capital letters (A) and the individual *entries* of the matrix, the numbers that make it up, will be denoted using lowercase letters (a_{ij}) where the first number i indicates which row of the matrix we are talking about, and the second number j indicates

which column. For example, in the matrix

$$A = \begin{bmatrix} 1 & 4 & 0 \\ -2 & 3 & 1 \\ 2 & 0 & 5 \end{bmatrix},$$

we could talk about the entire matrix usint A , but would also have that $a_{21} = -2$ and $a_{33} = 5$.

Note that an $m \times 1$ matrix is just a column vector, so in terms of the basic structure, matrices are an extension of vectors. However, they can be used for so much more, as we will see in future sections.

Another way to view matrices is as a set of column vectors all laid out side-by-side. If we have \vec{v}_1 , \vec{v}_2 and \vec{v}_3 , three different four component vectors, we can form a 4×3 matrix B as

$$B = [\vec{v}_1 \mid \vec{v}_2 \mid \vec{v}_3]$$

that uses each of the given vectors as a column of the matrix. In this case, the vertical lines are used to indicate that this is actually a matrix, because each of the entries given there are vectors, not just individual numbers. If we wanted to write a 1×3 matrix this way, these vertical lines will not be included.

We will go into more properties of matrices and the operations we can perform on them in § 3.2. To conclude this section though, we will look at one other way that matrices come about, and that is as the representation of a linear map.

3.1.3 Linear mappings and matrices

A *vector-valued function* F is a rule that takes a vector \vec{x} and returns another vector \vec{y} . For example, F could be a scaling that doubles the size of vectors:

$$F(\vec{x}) = 2\vec{x}.$$

For example,

$$F\left(\begin{bmatrix} 1 \\ 3 \end{bmatrix}\right) = 2 \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \end{bmatrix}.$$

If F is a mapping that takes vectors in \mathbb{R}^2 to \mathbb{R}^2 (such as the above), we write

$$F: \mathbb{R}^2 \rightarrow \mathbb{R}^2.$$

The words *function* and *mapping* are used rather interchangeably, although more often than not, *mapping* is used when talking about a vector-valued function, and the word *function* is often used when the function is scalar-valued.

A beginning student of mathematics (and many a seasoned mathematician), that sees an expression such as

$$f(3x + 8y)$$

yearns to write

$$3f(x) + 8f(y).$$

After all, who hasn't wanted to write $\sqrt{x+y} = \sqrt{x} + \sqrt{y}$ or something like that at some point in their mathematical lives. Wouldn't life be simple if we could do that? Of course we can't always do that (for example, not with the square roots!) It turns out there are many functions where we can do exactly the above. Such functions are called *linear*.

A mapping $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is called *linear* if

$$F(\vec{x} + \vec{y}) = F(\vec{x}) + F(\vec{y}),$$

for any vectors \vec{x} and \vec{y} , and also

$$F(\alpha\vec{x}) = \alpha F(\vec{x}),$$

for any scalar α . The F we defined above that doubles the size of all vectors is linear. Let us check:

$$F(\vec{x} + \vec{y}) = 2(\vec{x} + \vec{y}) = 2\vec{x} + 2\vec{y} = F(\vec{x}) + F(\vec{y}),$$

and also

$$F(\alpha\vec{x}) = 2\alpha\vec{x} = \alpha 2\vec{x} = \alpha F(\vec{x}).$$

We also call a linear function a *linear transformation*. If you want to be really fancy and impress your friends, you can call it a *linear operator*.

When a mapping is linear we often do not write the parentheses. We write simply

$$F\vec{x}$$

instead of $F(\vec{x})$. We do this because linearity means that the mapping F behaves like multiplying \vec{x} by "something." That something is a matrix.

A *matrix* is an $m \times n$ array of numbers (m rows and n columns). A 3×5 matrix is

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \end{bmatrix}.$$

The numbers a_{ij} are called *elements* or *entries*.

A column vector is simply an $m \times 1$ matrix. Similarly to a column vector there is also a *row vector*, which is a $1 \times n$ matrix. If we have an $n \times n$ matrix, where the number of rows is the same as the number of columns, then we say that it is a *square matrix*.

Now how does a matrix A relate to a linear mapping? Well a matrix tells you where certain special vectors go. Let's give a name to those certain vectors. The *standard basis vectors* of \mathbb{R}^n are

$$\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \vec{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \vec{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \quad \cdots, \quad \vec{e}_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$

For example, in \mathbb{R}^3 these vectors are

$$\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \vec{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

You may recall from calculus of several variables that these are sometimes called \vec{i} , \vec{j} , \vec{k} .

The reason these are called a *basis* is that every other vector can be written as a *linear combination* of them. For example, in \mathbb{R}^3 the vector $(4, 5, 6)$ can be written as

$$4\vec{e}_1 + 5\vec{e}_2 + 6\vec{e}_3 = 4 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 5 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 6 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}.$$

So how does a matrix represent a linear mapping? Well, the columns of the matrix are the vectors where A as a linear mapping takes \vec{e}_1 , \vec{e}_2 , etc. For example, consider

$$M = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

As a linear mapping $M: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ takes $\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ to $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ and $\vec{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ to $\begin{bmatrix} 2 \\ 4 \end{bmatrix}$. In other words,

$$M\vec{e}_1 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \quad \text{and} \quad M\vec{e}_2 = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}.$$

More generally, if we have an $n \times m$ matrix A , that is we have n rows and m columns, then the mapping $A: \mathbb{R}^m \rightarrow \mathbb{R}^n$ takes \vec{e}_j to the j^{th} column of A . For example,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \end{bmatrix}$$

represents a mapping from \mathbb{R}^5 to \mathbb{R}^3 that does

$$A\vec{e}_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix}, \quad A\vec{e}_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}, \quad A\vec{e}_3 = \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix}, \quad A\vec{e}_4 = \begin{bmatrix} a_{14} \\ a_{24} \\ a_{34} \end{bmatrix}, \quad A\vec{e}_5 = \begin{bmatrix} a_{15} \\ a_{25} \\ a_{35} \end{bmatrix}.$$

But what if I have another vector \vec{x} ? Where does it go? Well we use linearity. First write the vector as a linear combination of the standard basis vectors:

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + x_5 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3 + x_4\vec{e}_4 + x_5\vec{e}_5.$$

Then

$$A\vec{x} = A(x_1\vec{e}_1 + x_2\vec{e}_2 + x_3\vec{e}_3 + x_4\vec{e}_4 + x_5\vec{e}_5) = x_1A\vec{e}_1 + x_2A\vec{e}_2 + x_3A\vec{e}_3 + x_4A\vec{e}_4 + x_5A\vec{e}_5.$$

If we know where A takes all the basis vectors, we know where it takes all vectors.

As an example, suppose M is the 2×2 matrix from above, and suppose we wish to find

$$M \begin{bmatrix} -2 \\ 0.1 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} -2 \\ 0.1 \end{bmatrix} = -2 \begin{bmatrix} 1 \\ 3 \end{bmatrix} + 0.1 \begin{bmatrix} 2 \\ 4 \end{bmatrix} = \begin{bmatrix} -1.8 \\ -5.6 \end{bmatrix}.$$

Every linear mapping from \mathbb{R}^m to \mathbb{R}^n can be represented by an $n \times m$ matrix. You just figure out where it takes the standard basis vectors. Conversely, every $n \times m$ matrix represents a linear mapping. Hence, we may think of matrices being linear mappings, and linear mappings being matrices.

Or can we? In this book we study mostly linear differential operators, and linear differential operators are linear mappings, although they are not acting on \mathbb{R}^n , but on an infinite-dimensional space of functions:

$$Lf = g$$

for a function f we get a function g , and L is linear in the sense that

$$L(f + h) = Lf + Lh, \quad \text{and} \quad L(\alpha f) = \alpha Lf.$$

for any number (scalars) α and all functions f and h .

So the answer is not really. But if we consider vectors in finite-dimensional spaces \mathbb{R}^n then yes, every linear mapping is a matrix. We have mentioned at the beginning of this section, that we can “make everything a vector.” That’s not strictly true, but it is true approximately. Those “infinite-dimensional” spaces of functions can be approximated by a finite-dimensional space, and then linear operators are just matrices. So approximately, this is true. And as far as actual computations that we can do on a computer, we can work only with finitely many dimensions anyway. If you ask a computer or your calculator to plot a function, it samples the function at finitely many points and then connects the dots*. It does not actually give you infinitely many values. So the way that you have been using the computer or your calculator so far has already been a certain approximation of the space of functions by a finite-dimensional space.

3.1.4 Exercises

Exercise 3.1.1: On a piece of graph paper draw the vectors:

a) $\begin{bmatrix} 2 \\ 5 \end{bmatrix}$

b) $\begin{bmatrix} -2 \\ -4 \end{bmatrix}$

c) $(3, -4)$

Exercise 3.1.2: On a piece of graph paper draw the vector $(1, 2)$ starting at (based at) the given point:

a) based at $(0, 0)$

b) based at $(1, 2)$

c) based at $(0, -1)$

*In Matlab, you may have noticed that to plot a function, we take a vector of inputs, ask Matlab to compute the corresponding vector of values of the function, and then we ask it to plot the result.

Exercise 3.1.3: On a piece of graph paper draw the following operations. Draw and label the vectors involved in the operations as well as the result:

$$a) \begin{bmatrix} 1 \\ -4 \end{bmatrix} + \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$b) \begin{bmatrix} -3 \\ 2 \end{bmatrix} - \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$c) 3 \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Exercise 3.1.4: Compute the magnitude of

$$a) \begin{bmatrix} 7 \\ 2 \end{bmatrix}$$

$$b) \begin{bmatrix} -2 \\ 3 \\ 1 \end{bmatrix}$$

$$c) (1, 3, -4)$$

Exercise 3.1.5:* Compute the magnitude of

$$a) \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$b) \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}$$

$$c) (-2, 1, -2)$$

Exercise 3.1.6: Compute

$$a) \begin{bmatrix} 2 \\ 3 \end{bmatrix} + \begin{bmatrix} 7 \\ -8 \end{bmatrix}$$

$$b) \begin{bmatrix} -2 \\ 3 \end{bmatrix} - \begin{bmatrix} 6 \\ -4 \end{bmatrix}$$

$$c) - \begin{bmatrix} -3 \\ 2 \end{bmatrix}$$

$$d) 4 \begin{bmatrix} -1 \\ 5 \end{bmatrix}$$

$$e) 5 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 9 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$f) 3 \begin{bmatrix} 1 \\ -8 \end{bmatrix} - 2 \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

Exercise 3.1.7:* Compute

$$a) \begin{bmatrix} 3 \\ 1 \end{bmatrix} + \begin{bmatrix} 6 \\ -3 \end{bmatrix}$$

$$b) \begin{bmatrix} -1 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

$$c) - \begin{bmatrix} -5 \\ 3 \end{bmatrix}$$

$$d) 2 \begin{bmatrix} -2 \\ 4 \end{bmatrix}$$

$$e) 3 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 7 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$f) 2 \begin{bmatrix} 2 \\ -3 \end{bmatrix} - 6 \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

Exercise 3.1.8: Find the unit vector in the direction of the given vector

$$a) \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

$$b) \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$$

$$c) (3, 1, -2)$$

Exercise 3.1.9:* Find the unit vector in the direction of the given vector

$$a) \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$b) \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$$

$$c) (2, -5, 2)$$

Exercise 3.1.10: If $\vec{x} = (1, 2)$ and \vec{y} are added together, we find $\vec{x} + \vec{y} = (0, 2)$. What is \vec{y} ?

Exercise 3.1.11: If $\vec{v} = (1, -4, 3)$ and $\vec{w} = (-2, 3, -1)$, compute $3\vec{v} - 2\vec{w}$ and $4\vec{w} + \vec{v}$.

Exercise 3.1.12: Write $(1, 2, 3)$ as a linear combination of the standard basis vectors \vec{e}_1 , \vec{e}_2 , and \vec{e}_3 .

Exercise 3.1.13: If the magnitude of \vec{x} is 4, what is the magnitude of

- a) $0\vec{x}$ b) $3\vec{x}$ c) $-\vec{x}$ d) $-4\vec{x}$ e) $\vec{x} + \vec{x}$ f) $\vec{x} - \vec{x}$

Exercise 3.1.14:* If the magnitude of \vec{x} is 5, what is the magnitude of

- a) $4\vec{x}$ b) $-2\vec{x}$ c) $-4\vec{x}$

Exercise 3.1.15: Suppose a linear mapping $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ takes $(1, 0)$ to $(2, -1)$ and it takes $(0, 1)$ to $(3, 3)$. Where does it take

- a) $(1, 1)$ b) $(2, 0)$ c) $(2, -1)$

Exercise 3.1.16: Suppose a linear mapping $F: \mathbb{R}^3 \rightarrow \mathbb{R}^2$ takes $(1, 0, 0)$ to $(2, 1)$ and it takes $(0, 1, 0)$ to $(3, 4)$ and it takes $(0, 0, 1)$ to $(5, 6)$. Write down the matrix representing the mapping F .

Exercise 3.1.17: Suppose that a mapping $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ takes $(1, 0)$ to $(1, 2)$, $(0, 1)$ to $(3, 4)$, and it takes $(1, 1)$ to $(0, -1)$. Explain why F is not linear.

Exercise 3.1.18:* Suppose a linear mapping $F: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ takes $(1, 0)$ to $(1, -1)$ and it takes $(0, 1)$ to $(2, 0)$. Where does it take

- a) $(1, 1)$ b) $(0, 2)$ c) $(1, -1)$

Exercise 3.1.19 (challenging): Let \mathbb{R}^3 represent the space of quadratic polynomials in t : a point (a_0, a_1, a_2) in \mathbb{R}^3 represents the polynomial $a_0 + a_1t + a_2t^2$. Consider the derivative $\frac{d}{dt}$ as a mapping of \mathbb{R}^3 to \mathbb{R}^3 , and note that $\frac{d}{dt}$ is linear. Write down $\frac{d}{dt}$ as a 3×3 matrix.

3.2 Matrix algebra

Attribution: [JL], §A.2.

Learning Objectives

After this section, you will be able to:

- Perform addition and multiplication operations on matrices,
- Compute inverses of 2×2 matrices, and
- Identify triangular, diagonal, and symmetric matrices.

3.2.1 One-by-one matrices

Let us motivate what we want to achieve with matrices. What do real-valued linear mappings look like? A linear function of real numbers that you have seen in calculus is of the form

$$f(x) = mx + b.$$

However, the properties of linear mappings discussed in the previous section are that

$$f(x+y) = f(x) + f(y) \quad f(ax) = af(x).$$

Plugging in the definition from above gives that

$$\begin{aligned} f(x+y) &= m(x+y) + b = mx + my + b \\ f(ax) &= m(ax) + b = a(mx) + b \end{aligned}$$

and neither of these match up appropriately, since

$$\begin{aligned} f(x) + f(y) &= mx + b + my + b = mx + my + 2b \\ af(x) + a(mx+b) &= a(mx) + ab \end{aligned}.$$

In order for these to work, we need to have $b = 0$. Therefore, real-valued linear mappings of the real line, linear functions that eat numbers and spit out numbers, are just multiplications by a number.

Consider a mapping defined by multiplying by a number. Let's call this number α . The mapping then takes x to αx . What we can do is to *add* such mappings. If we have another mapping β , then

$$\alpha x + \beta x = (\alpha + \beta)x.$$

We get a new mapping $\alpha + \beta$ that multiplies x by, well, $\alpha + \beta$. If D is a mapping that doubles things, $Dx = 2x$, and T is a mapping that triples, $Tx = 3x$, then $D + T$ is a mapping that multiplies by 5, $(D + T)x = 5x$.

Similarly we can *compose* such mappings, that is, we could apply one and then the other. We take x , we run it through the first mapping α to get α times x , then we run αx through the second mapping β . In other words,

$$\beta(\alpha x) = (\beta\alpha)x.$$

We just multiply those two numbers. Using our doubling and tripling mappings, if we double and then triple, that is $T(Dx)$ then we obtain $3(2x) = 6x$. The composition TD is the mapping that multiplies by 6. For larger matrices, composition also ends up being a kind of multiplication.

3.2.2 Matrix addition and scalar multiplication

The mappings that multiply numbers by numbers are just 1×1 matrices. The number α above could be written as a matrix $[\alpha]$. So perhaps we would want to do the same things to all matrices that we did to those 1×1 matrices at the start of this section above. First, let us add matrices. If we have a matrix A and a matrix B that are of the same size, say $m \times n$, then they are mappings from \mathbb{R}^n to \mathbb{R}^m . The mapping $A + B$ should also be a mapping from \mathbb{R}^n to \mathbb{R}^m , and it should do the following to vectors:

$$(A + B)\vec{x} = A\vec{x} + B\vec{x}.$$

It turns out you just add the matrices element-wise: If the ij^{th} entry of A is a_{ij} , and the ij^{th} entry of B is b_{ij} , then the ij^{th} entry of $A + B$ is $a_{ij} + b_{ij}$. If

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix},$$

then

$$A + B = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & a_{13} + b_{13} \\ a_{21} + b_{21} & a_{22} + b_{22} & a_{23} + b_{23} \end{bmatrix}.$$

Let us illustrate on a more concrete example:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} + \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & -1 \end{bmatrix} = \begin{bmatrix} 1+7 & 2+8 \\ 3+9 & 4+10 \\ 5+11 & 6-1 \end{bmatrix} = \begin{bmatrix} 8 & 10 \\ 12 & 14 \\ 16 & 5 \end{bmatrix}.$$

Let's check that this does the right thing to a vector. Let's use some of the vector algebra that we already know, and regroup things:

$$\begin{aligned} \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} + \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} &= \left(2 \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} - \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} \right) + \left(2 \begin{bmatrix} 7 \\ 9 \\ 11 \end{bmatrix} - \begin{bmatrix} 8 \\ 10 \\ -1 \end{bmatrix} \right) \\ &= 2 \left(\begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} + \begin{bmatrix} 7 \\ 9 \\ 11 \end{bmatrix} \right) - \left(\begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix} + \begin{bmatrix} 8 \\ 10 \\ -1 \end{bmatrix} \right) \\ &= 2 \begin{bmatrix} 1+7 \\ 3+9 \\ 5+11 \end{bmatrix} - \begin{bmatrix} 2+8 \\ 4+10 \\ 6-1 \end{bmatrix} = 2 \begin{bmatrix} 8 \\ 12 \\ 16 \end{bmatrix} - \begin{bmatrix} 10 \\ 14 \\ 5 \end{bmatrix} \\ &= \begin{bmatrix} 8 & 10 \\ 12 & 14 \\ 16 & 5 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad \left(= \begin{bmatrix} 2(8)-10 \\ 2(12)-14 \\ 2(16)-5 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 27 \end{bmatrix} \right). \end{aligned}$$

If we replaced the numbers by letters that would constitute a proof! You'll notice that we didn't really have to even compute what the result is to convince ourselves that the two expressions were equal.

If the sizes of the matrices do not match, then addition is not defined. If A is 3×2 and B is 2×5 , then we cannot add these matrices. We don't know what that could possibly mean.

It is also useful to have a matrix that when added to any other matrix does nothing. This is the zero matrix, the matrix of all zeros:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

We often denote the zero matrix by 0 without specifying size. We would then just write $A + 0$, where we just assume that 0 is the zero matrix of the same size as A .

There are really two things we can multiply matrices by. We can multiply matrices by scalars or we can multiply by other matrices. Let us first consider multiplication by scalars. For a matrix A and a scalar α we want αA to be the matrix that accomplishes

$$(\alpha A)\vec{x} = \alpha(A\vec{x}).$$

That is just scaling the result by α . If you think about it, scaling every term in A by α accomplishes just that: If

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}, \quad \text{then} \quad \alpha A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \alpha a_{13} \\ \alpha a_{21} & \alpha a_{22} & \alpha a_{23} \end{bmatrix}.$$

For example,

$$2 \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{bmatrix}.$$

Let us list some properties of matrix addition and scalar multiplication. Denote by 0 the zero matrix, by α, β scalars, and by A, B, C matrices. Then:

$$\begin{aligned} A + 0 &= A = 0 + A, \\ A + B &= B + A, \\ (A + B) + C &= A + (B + C), \\ \alpha(A + B) &= \alpha A + \alpha B, \\ (\alpha + \beta)A &= \alpha A + \beta A. \end{aligned}$$

These rules should look very familiar.

3.2.3 Matrix multiplication

As we mentioned above, composition of linear mappings is also a multiplication of matrices. Suppose A is an $m \times n$ matrix, that is, A takes \mathbb{R}^n to \mathbb{R}^m , and B is an $n \times p$ matrix, that is, B takes \mathbb{R}^p to \mathbb{R}^n . The composition AB should work as follows

$$AB\vec{x} = A(B\vec{x}).$$

First, a vector \vec{x} in \mathbb{R}^p gets taken to the vector $B\vec{x}$ in \mathbb{R}^n . Then the mapping A takes it to the vector $A(B\vec{x})$ in \mathbb{R}^m . In other words, the composition AB should be an $m \times p$ matrix. In terms of sizes we should have

$$\text{“ } [m \times n] [n \times p] = [m \times p]. \text{ ”}$$

Notice how the middle size must match.

OK, now we know what sizes of matrices we should be able to multiply, and what the product should be. Let us see how to actually compute matrix multiplication. We start with the so-called *dot product* (or *inner product*) of two vectors. Usually this is a row vector multiplied with a column vector of the same size. Dot product multiplies each pair of entries from the first and the second vector and sums these products. The result is a single number. For example,

$$\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3.$$

And similarly for larger (or smaller) vectors. A dot product is really a product of two matrices: a $1 \times n$ matrix and an $n \times 1$ matrix resulting in a 1×1 matrix, that is, a number.

Armed with the dot product we define the *product of matrices*. First let us denote by $\text{row}_i(A)$ the i^{th} row of A and by $\text{column}_j(A)$ the j^{th} column of A . For an $m \times n$ matrix A and an $n \times p$ matrix B we can compute the product AB . The matrix AB is an $m \times p$ matrix whose ij^{th} entry is the dot product

$$\text{row}_i(A) \cdot \text{column}_j(B).$$

For example, given a 2×3 and a 3×2 matrix we should end up with a 2×2 matrix:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix}, \quad (3.1)$$

or with some numbers:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} -1 & 2 \\ -7 & 0 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 \cdot (-1) + 2 \cdot (-7) + 3 \cdot 1 & 1 \cdot 2 + 2 \cdot 0 + 3 \cdot (-1) \\ 4 \cdot (-1) + 5 \cdot (-7) + 6 \cdot 1 & 4 \cdot 2 + 5 \cdot 0 + 6 \cdot (-1) \end{bmatrix} = \begin{bmatrix} -12 & -1 \\ -33 & 2 \end{bmatrix}.$$

A useful consequence of the definition is that the evaluation $A\vec{x}$ for a matrix A and a (column) vector \vec{x} is also matrix multiplication. That is really why we think of vectors as column vectors, or $n \times 1$ matrices. For example,

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 2 + 2 \cdot (-1) \\ 3 \cdot 2 + 4 \cdot (-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

If you look at the last section, that is precisely the last example we gave.

You should stare at the computation of multiplication of matrices AB and the previous definition of $A\vec{y}$ as a mapping for a moment. What we are doing with matrix multiplication

is applying the mapping A to the columns of B . This is usually written as follows. Suppose we write the $n \times p$ matrix $B = [\vec{b}_1 \ \vec{b}_2 \ \dots \ \vec{b}_p]$, where $\vec{b}_1, \vec{b}_2, \dots, \vec{b}_p$ are the columns of B . Then for an $m \times n$ matrix A ,

$$AB = A[\vec{b}_1 \ \vec{b}_2 \ \dots \ \vec{b}_p] = [A\vec{b}_1 \ A\vec{b}_2 \ \dots \ A\vec{b}_p].$$

The columns of the $m \times p$ matrix AB are the vectors $A\vec{b}_1, A\vec{b}_2, \dots, A\vec{b}_p$. For example, in (3.1), the columns of

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}$$

are

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{21} \\ b_{31} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{22} \\ b_{32} \end{bmatrix}.$$

This is a very useful way to understand what matrix multiplication is. It should also make it easier to remember how to perform matrix multiplication.

3.2.4 Some rules of matrix algebra

For multiplication we want an analogue of a 1. That is, we desire a matrix that just leaves everything as it found it. This analogue is the so-called *identity matrix*. The identity matrix is a square matrix with 1s on the main diagonal and zeros everywhere else. It is usually denoted by I . For each size we have a different identity matrix and so sometimes we may denote the size as a subscript. For example, the I_3 would be the 3×3 identity matrix

$$I = I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let us see how the matrix works on a smaller example,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} a_{11} \cdot 1 + a_{12} \cdot 0 & a_{11} \cdot 0 + a_{12} \cdot 1 \\ a_{21} \cdot 1 + a_{22} \cdot 0 & a_{21} \cdot 0 + a_{22} \cdot 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Multiplication by the identity from the left looks similar, and also does not touch anything.

We have the following rules for matrix multiplication. Suppose that A, B, C are matrices of the correct sizes so that the following make sense. Let α denote a scalar (number). Then

$$\begin{aligned} A(BC) &= (AB)C && \text{(associative law),} \\ A(B+C) &= AB+AC && \text{(distributive law),} \\ (B+C)A &= BA+CA && \text{(distributive law),} \\ \alpha(AB) &= (\alpha A)B = A(\alpha B), \\ IA &= A = AI && \text{(identity).} \end{aligned}$$

Example 3.2.1: Let us demonstrate a couple of these rules. For example, the associative law:

$$\underbrace{\begin{bmatrix} -3 & 3 \\ 2 & -2 \end{bmatrix}}_A \left(\underbrace{\begin{bmatrix} 4 & 4 \\ 1 & -3 \end{bmatrix}}_B \underbrace{\begin{bmatrix} -1 & 4 \\ 5 & 2 \end{bmatrix}}_C \right) = \underbrace{\begin{bmatrix} -3 & 3 \\ 2 & -2 \end{bmatrix}}_A \underbrace{\begin{bmatrix} 16 & 24 \\ -16 & -2 \end{bmatrix}}_{BC} = \underbrace{\begin{bmatrix} -96 & -78 \\ 64 & 52 \end{bmatrix}}_{A(BC)},$$

and

$$\left(\underbrace{\begin{bmatrix} -3 & 3 \\ 2 & -2 \end{bmatrix}}_A \underbrace{\begin{bmatrix} 4 & 4 \\ 1 & -3 \end{bmatrix}}_B \right) \underbrace{\begin{bmatrix} -1 & 4 \\ 5 & 2 \end{bmatrix}}_C = \underbrace{\begin{bmatrix} -9 & -21 \\ 6 & 14 \end{bmatrix}}_{AB} \underbrace{\begin{bmatrix} -1 & 4 \\ 5 & 2 \end{bmatrix}}_C = \underbrace{\begin{bmatrix} -96 & -78 \\ 64 & 52 \end{bmatrix}}_{(AB)C}.$$

Or how about multiplication by scalars:

$$10 \left(\underbrace{\begin{bmatrix} -3 & 3 \\ 2 & -2 \end{bmatrix}}_A \underbrace{\begin{bmatrix} 4 & 4 \\ 1 & -3 \end{bmatrix}}_B \right) = 10 \underbrace{\begin{bmatrix} -9 & -21 \\ 6 & 14 \end{bmatrix}}_{AB} = \underbrace{\begin{bmatrix} -90 & -210 \\ 60 & 140 \end{bmatrix}}_{10(AB)},$$

$$\left(10 \underbrace{\begin{bmatrix} -3 & 3 \\ 2 & -2 \end{bmatrix}}_A \right) \underbrace{\begin{bmatrix} 4 & 4 \\ 1 & -3 \end{bmatrix}}_B = \underbrace{\begin{bmatrix} -30 & 30 \\ 20 & -20 \end{bmatrix}}_{10A} \underbrace{\begin{bmatrix} 4 & 4 \\ 1 & -3 \end{bmatrix}}_B = \underbrace{\begin{bmatrix} -90 & -210 \\ 60 & 140 \end{bmatrix}}_{(10A)B},$$

and

$$\underbrace{\begin{bmatrix} -3 & 3 \\ 2 & -2 \end{bmatrix}}_A \left(10 \underbrace{\begin{bmatrix} 4 & 4 \\ 1 & -3 \end{bmatrix}}_B \right) = \underbrace{\begin{bmatrix} -3 & 3 \\ 2 & -2 \end{bmatrix}}_A \underbrace{\begin{bmatrix} 40 & 40 \\ 10 & -30 \end{bmatrix}}_{10B} = \underbrace{\begin{bmatrix} -90 & -210 \\ 60 & 140 \end{bmatrix}}_{A(10B)}.$$

A multiplication rule you have used since primary school on numbers is quite conspicuously missing for matrices. That is, matrix multiplication is not commutative. Firstly, just because AB makes sense, it may be that BA is not even defined. For example, if A is 2×3 , and B is 3×4 , then we can multiply AB but not BA .

Even if AB and BA are both defined, does not mean that they are equal. For example, take $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$:

$$AB = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} \neq \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = BA.$$

3.2.5 Inverse

A couple of other algebra rules you know for numbers do not quite work on matrices:

- (i) $AB = AC$ does not necessarily imply $B = C$, even if A is not 0.
- (ii) $AB = 0$ does not necessarily mean that $A = 0$ or $B = 0$.

For example:

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}.$$

To make these rules hold, we do not just need one of the matrices to not be zero, we would need to “divide” by a matrix. This is where the *matrix inverse* comes in.

Definition 3.2.1

Suppose that A and B are $n \times n$ matrices such that

$$AB = I = BA.$$

Then we call B the inverse of A and we denote B by A^{-1} .

If the inverse of A exists, then we say A is *invertible*. If A is not invertible, we say A is *singular*.

Perhaps not surprisingly, $(A^{-1})^{-1} = A$, since if the inverse of A is B , then the inverse of B is A .

If $A = [a]$ is a 1×1 matrix, then A^{-1} is $a^{-1} = \frac{1}{a}$. That is where the notation comes from. The computation is not nearly as simple when A is larger.

The proper formulation of the cancellation rule is:

If A is invertible, then $AB = AC$ implies $B = C$.

The computation is what you would do in regular algebra with numbers, but you have to be careful never to commute matrices:

$$\begin{aligned} AB &= AC, \\ A^{-1}AB &= A^{-1}AC, \\ IB &= IC, \\ B &= C. \end{aligned}$$

And similarly for cancellation on the right:

If A is invertible, then $BA = CA$ implies $B = C$.

The rule says, among other things, that the inverse of a matrix is unique if it exists: If $AB = I = AC$, then A is invertible and $B = C$.

We will see later how to compute an inverse of a matrix in general. For now, let us note that there is a simple formula for the inverse of a 2×2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

For example:

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix}^{-1} = \frac{1}{1 \cdot 4 - 1 \cdot 2} \begin{bmatrix} 4 & -1 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1/2 \\ -1 & 1/2 \end{bmatrix}.$$

Let's try it:

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 2 & -1/2 \\ -1 & 1/2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 & -1/2 \\ -1 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Just as we cannot divide by every number, not every matrix is invertible. In the case of matrices however we may have singular matrices that are not zero. For example,

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$$

is a singular matrix. But didn't we just give a formula for an inverse? Let us try it:

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}^{-1} = \frac{1}{1 \cdot 2 - 1 \cdot 2} \begin{bmatrix} 2 & -1 \\ -2 & 1 \end{bmatrix} = ?$$

We get into a bit of trouble; we are trying to divide by zero.

So a 2×2 matrix A is invertible whenever

$$ad - bc \neq 0$$

and otherwise it is singular. The expression $ad - bc$ is called the *determinant* and we will look at it more carefully in a later section. There is a similar expression for a square matrix of any size.

3.2.6 Triangular and Diagonal matrices

A simple (and surprisingly useful) type of a square matrix is a so-called *diagonal matrix*. It is a matrix whose entries are all zero except those on the main diagonal from top left to bottom right. For example a 4×4 diagonal matrix is of the form

$$\begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & d_3 & 0 \\ 0 & 0 & 0 & d_4 \end{bmatrix}.$$

Such matrices have nice properties when we multiply by them. If we multiply them by a vector, they multiply the k^{th} entry by d_k . For example,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 \cdot 4 \\ 2 \cdot 5 \\ 3 \cdot 6 \end{bmatrix} = \begin{bmatrix} 4 \\ 10 \\ 18 \end{bmatrix}.$$

Similarly, when they multiply another matrix from the left, they multiply the k^{th} row by d_k . For example,

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 2 \\ 3 & 3 & 3 \\ -1 & -1 & -1 \end{bmatrix}.$$

On the other hand, multiplying on the right, they multiply the columns:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 3 & -1 \\ 2 & 3 & -1 \\ 2 & 3 & -1 \end{bmatrix}.$$

And it is really easy to multiply two diagonal matrices together:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 2 & 0 & 0 \\ 0 & 2 \cdot 3 & 0 \\ 0 & 0 & 3 \cdot (-1) \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & -3 \end{bmatrix}.$$

For this last reason, they are easy to invert, you simply invert each diagonal element:

$$\begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix}^{-1} = \begin{bmatrix} d_1^{-1} & 0 & 0 \\ 0 & d_2^{-1} & 0 \\ 0 & 0 & d_3^{-1} \end{bmatrix}.$$

Let us check an example

$$\underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}}_{A^{-1}} \underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}}_A = \underbrace{\begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}}_{A^{-1}} \underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}}_A = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_I.$$

It is no wonder that the way we solve many problems in linear algebra (and in differential equations) is to try to reduce the problem to the case of diagonal matrices.

Another type of matrix that has similarly nice properties are *triangular* matrices. A matrix is *upper triangular* if all of the entries below the diagonal are zero. For a 3×3 matrix, an upper triangular matrix looks like

$$\begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix}$$

where the $*$ can be any number. Similarly, a *lower triangular* matrix is one where all of the entries above the diagonal are zero, or, for a 3×3 matrix, something that looks like

$$\begin{bmatrix} * & 0 & 0 \\ * & * & 0 \\ * & * & * \end{bmatrix}.$$

A matrix that is both upper and lower triangular is diagonal, because only the entries on the diagonal can be non-zero.

3.2.7 Transpose

Vectors do not always have to be column vectors, that is just a convention. Swapping rows and columns is from time to time needed. The operation that swaps rows and columns is the so-called *transpose*. The transpose of A is denoted by A^T . Example:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

So transpose takes an $m \times n$ matrix to an $n \times m$ matrix.

A key fact about the transpose is that if the product AB makes sense then $B^T A^T$ also makes sense, at least from the point of view of sizes. In fact, we get precisely the transpose of AB . That is:

$$(AB)^T = B^T A^T.$$

For example,

$$\left(\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & -2 \end{bmatrix} \right)^T = \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

It is left to the reader to verify that computing the matrix product on the left and then transposing is the same as computing the matrix product on the right.

If we have a column vector \vec{x} to which we apply a matrix A and we transpose the result, then the row vector \vec{x}^T applies to A^T from the left:

$$(A\vec{x})^T = \vec{x}^T A^T.$$

Another place where transpose is useful is when we wish to apply the dot product* to two column vectors:

$$\vec{x} \cdot \vec{y} = \vec{y}^T \vec{x}.$$

That is the way that one often writes the dot product in software.

We say a matrix A is *symmetric* if $A = A^T$. For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

is a symmetric matrix. Notice that a symmetric matrix is always square, that is, $n \times n$. Symmetric matrices have many nice properties†, and come up quite often in applications.

To end the section, we notice how $A\vec{x}$ can be written more succinctly. Suppose

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \quad \text{and} \quad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Then

$$A\vec{x} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{bmatrix}.$$

For example,

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 2 + 2 \cdot (-1) \\ 3 \cdot 2 + 4 \cdot (-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

In other words, you take a row of the matrix, you multiply them by the entries in your vector, you add things up, and that's the corresponding entry in the resulting vector.

*As a side note, mathematicians write $\vec{y}^T \vec{x}$ and physicists write $\vec{x}^T \vec{y}$. Shhh... don't tell anyone, but the physicists are probably right on this.

†Although so far we have not learned enough about matrices to really appreciate them.

3.2.8 Exercises

Exercise 3.2.1: Add the following matrices

$$a) \begin{bmatrix} -1 & 2 & 2 \\ 5 & 8 & -1 \end{bmatrix} + \begin{bmatrix} 3 & 2 & 3 \\ 8 & 3 & 5 \end{bmatrix}$$

$$b) \begin{bmatrix} 1 & 2 & 4 \\ 2 & 3 & 1 \\ 0 & 5 & 1 \end{bmatrix} + \begin{bmatrix} 2 & -8 & -3 \\ 3 & 1 & 0 \\ 6 & -4 & 1 \end{bmatrix}$$

Exercise 3.2.2:* Add the following matrices

$$a) \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & -1 \end{bmatrix} + \begin{bmatrix} 5 & 3 & 4 \\ 1 & 2 & 5 \end{bmatrix}$$

$$b) \begin{bmatrix} 6 & -2 & 3 \\ 7 & 3 & 3 \\ 8 & -1 & 2 \end{bmatrix} + \begin{bmatrix} -1 & -1 & -3 \\ 6 & 7 & 3 \\ -9 & 4 & -1 \end{bmatrix}$$

Exercise 3.2.3: Compute

$$a) 3 \begin{bmatrix} 0 & 3 \\ -2 & 2 \end{bmatrix} + 6 \begin{bmatrix} 1 & 5 \\ -1 & 5 \end{bmatrix}$$

$$b) 2 \begin{bmatrix} -3 & 1 \\ 2 & 2 \end{bmatrix} - 3 \begin{bmatrix} 2 & -1 \\ 3 & 2 \end{bmatrix}$$

Exercise 3.2.4:* Compute

$$a) 2 \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + 3 \begin{bmatrix} -1 & 3 \\ 1 & 2 \end{bmatrix}$$

$$b) 3 \begin{bmatrix} 2 & -1 \\ 1 & 3 \end{bmatrix} - 2 \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}$$

Exercise 3.2.5: Multiply the following matrices

$$a) \begin{bmatrix} -1 & 2 \\ 3 & 1 \\ 5 & 8 \end{bmatrix} \begin{bmatrix} 3 & -1 & 3 & 1 \\ 8 & 3 & 2 & -3 \end{bmatrix}$$

$$b) \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 1 \\ 1 & 0 & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 & 1 & 7 \\ 1 & 2 & 3 & -1 \\ 1 & -1 & 3 & 0 \end{bmatrix}$$

$$c) \begin{bmatrix} 4 & 1 & 6 & 3 \\ 5 & 6 & 5 & 0 \\ 4 & 6 & 6 & 0 \end{bmatrix} \begin{bmatrix} 2 & 5 \\ 1 & 2 \\ 3 & 5 \\ 5 & 6 \end{bmatrix}$$

$$d) \begin{bmatrix} 1 & 1 & 4 \\ 0 & 5 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 1 & 0 \\ 6 & 4 \end{bmatrix}$$

Exercise 3.2.6:* Multiply the following matrices

$$a) \begin{bmatrix} 2 & 1 & 4 \\ 3 & 4 & 4 \end{bmatrix} \begin{bmatrix} 2 & 4 \\ 6 & 3 \\ 3 & 5 \end{bmatrix}$$

$$b) \begin{bmatrix} 0 & 3 & 3 \\ 2 & -2 & 1 \\ 3 & 5 & -2 \end{bmatrix} \begin{bmatrix} 6 & 6 & 2 \\ 4 & 6 & 0 \\ 2 & 0 & 4 \end{bmatrix}$$

$$c) \begin{bmatrix} 3 & 4 & 1 \\ 2 & -1 & 0 \\ 4 & -1 & 5 \end{bmatrix} \begin{bmatrix} 0 & 2 & 5 & 0 \\ 2 & 0 & 5 & 2 \\ 3 & 6 & 1 & 6 \end{bmatrix}$$

$$d) \begin{bmatrix} -2 & -2 \\ 5 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 0 & 3 \\ 1 & 3 \end{bmatrix}$$

Exercise 3.2.7:

- a) How must the dimensions of two matrices line up in order to multiply them together? If they can be multiplied, what is the dimension of the product?
- b) If A is a 3×2 matrix and the product AB is a 3×4 matrix, then what are the dimensions of B ?
- c) If A is a 5×3 matrix, is it possible to find a matrix B so that the product AB is a 4×3 matrix? What about a matrix C so that the product CA is a 4×3 matrix?

Exercise 3.2.8: Compute the inverse of the given matrices

a) $[-3]$

b) $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$

c) $\begin{bmatrix} 1 & 4 \\ 1 & 3 \end{bmatrix}$

d) $\begin{bmatrix} 2 & 2 \\ 1 & 4 \end{bmatrix}$

Exercise 3.2.9:* Compute the inverse of the given matrices

a) $[2]$

b) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

c) $\begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix}$

d) $\begin{bmatrix} 4 & 2 \\ 4 & 4 \end{bmatrix}$

Exercise 3.2.10: Compute the inverse of the given matrices

a) $\begin{bmatrix} -2 & 0 \\ 0 & 1 \end{bmatrix}$

b) $\begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

c) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & -5 \end{bmatrix}$

Exercise 3.2.11:* Compute the inverse of the given matrices

a) $\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$

b) $\begin{bmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & -1 \end{bmatrix}$

c) $\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix}$

3.3 Elimination

Attribution: [JL], §A.3.

Learning Objectives

After this section, you will be able to:

- Write a system of linear equations in matrix form,
- Use row reduction to put a matrix into row echelon form or reduced row echelon form,
- Determine whether a system of linear equations has no solution, one solution, or infinitely many solutions, and
- Compute the inverse of a matrix using row reduction.

3.3.1 Linear systems of equations

One application of matrices is to solve systems of linear equations*. Consider the following system of linear equations

$$\begin{aligned} 2x_1 + 2x_2 + 2x_3 &= 2, \\ x_1 + x_2 + 3x_3 &= 5, \\ x_1 + 4x_2 + x_3 &= 10. \end{aligned} \tag{3.2}$$

There is a systematic procedure called *elimination* to solve such a system. In this procedure, we attempt to eliminate each variable from all but one equation. We want to end up with equations such as $x_3 = 2$, where we can just read off the answer.

We write a system of linear equations as a matrix equation:

$$A\vec{x} = \vec{b}.$$

The system (3.2) is written as

$$\underbrace{\begin{bmatrix} 2 & 2 & 2 \\ 1 & 1 & 3 \\ 1 & 4 & 1 \end{bmatrix}}_A \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{\vec{x}} = \underbrace{\begin{bmatrix} 2 \\ 5 \\ 10 \end{bmatrix}}_{\vec{b}}.$$

If we knew the inverse of A , then we would be done; we would simply solve the equation:

$$\vec{x} = A^{-1}A\vec{x} = A^{-1}\vec{b}.$$

Well, but that is part of the problem, we do not know how to compute the inverse for matrices bigger than 2×2 . We will see later that to compute the inverse we are really solving $A\vec{x} = \vec{b}$

*Although perhaps we have this backwards, quite often we solve a linear system of equations to find out something about matrices, rather than vice versa.

for several different \vec{b} . In other words, we will need to do elimination to find A^{-1} . In addition, we may wish to solve $A\vec{x} = \vec{b}$ even if A is not invertible, or perhaps not even square.

Let us return to the equations themselves and see how we can manipulate them. There are a few operations we can perform on the equations that do not change the solution. First, perhaps an operation that may seem stupid, we can swap two equations in (3.2):

$$\begin{aligned}x_1 + x_2 + 3x_3 &= 5, \\2x_1 + 2x_2 + 2x_3 &= 2, \\x_1 + 4x_2 + x_3 &= 10.\end{aligned}$$

Clearly these new equations have the same solutions x_1, x_2, x_3 . A second operation is that we can multiply an equation by a nonzero number. For example, we multiply the third equation in (3.2) by 3:

$$\begin{aligned}2x_1 + 2x_2 + 2x_3 &= 2, \\x_1 + x_2 + 3x_3 &= 5, \\3x_1 + 12x_2 + 3x_3 &= 30.\end{aligned}$$

Finally we can add a multiple of one equation to another equation. For example, we add 3 times the third equation in (3.2) to the second equation:

$$\begin{aligned}2x_1 + 2x_2 + 2x_3 &= 2, \\(1+3)x_1 + (1+12)x_2 + (3+3)x_3 &= 5+30, \\x_1 + 4x_2 + x_3 &= 10.\end{aligned}$$

The same x_1, x_2, x_3 should still be solutions to the new equations. These were just examples; we did not get any closer to the solution. We must do these three operations in some more logical manner, but it turns out these three operations suffice to solve every linear equation.

The first thing is to write the equations in a more compact manner. Given

$$A\vec{x} = \vec{b},$$

we write down the so-called *augmented matrix*

$$[A \mid \vec{b}],$$

where the vertical line is just a marker for us to know where the “right-hand side” of the equation starts. For example, for the system (3.2) the augmented matrix is

$$\left[\begin{array}{ccc|c} 2 & 2 & 2 & 2 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{array} \right].$$

The entire process of elimination, which we will describe, is often applied to any sort of matrix, not just an augmented matrix. Simply think of the matrix as the 3×4 matrix

$$\left[\begin{array}{cccc} 2 & 2 & 2 & 2 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{array} \right].$$

3.3.2 Row echelon form and elementary operations

We apply the three operations above to the matrix. We call these the *elementary operations* or *elementary row operations*.

Definition 3.3.1

The elementary row operations on a matrix are:

- (i) Swap two rows.
- (ii) Multiply a row by a nonzero number.
- (iii) Add a multiple of one row to another row.

We run these operations until we get into a state where it is easy to read off the answer, or until we get into a contradiction indicating no solution.

More specifically, we run the operations until we obtain the so-called *row echelon form*. Let us call the first (from the left) nonzero entry in each row the *leading entry*. A matrix is in *row echelon form* if the following conditions are satisfied:

- (i) The leading entry in any row is strictly to the right of the leading entry of the row above.
- (ii) Any zero rows are below all the nonzero rows.
- (iii) All leading entries are 1.

A matrix is in *reduced row echelon form* if furthermore the following condition is satisfied.

- (iv) All the entries above a leading entry are zero.

Example 3.3.1: The following matrices are in row echelon form. The leading entries are marked:

$$\begin{bmatrix} 1 & 2 & 9 & 3 \\ 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & -1 & -3 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & -5 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Note that the definition applies to matrices of any size. None of the matrices above are in *reduced row echelon form*. For example, in the first matrix none of the entries above the second and third leading entries are zero; they are 9, 3, and 5.

The following matrices are in reduced row echelon form. The leading entries are marked:

$$\begin{bmatrix} 1 & 3 & 0 & 8 \\ 0 & 0 & 1 & 6 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

The procedure we will describe to find a reduced row echelon form of a matrix is called *Gauss–Jordan elimination*. The first part of it, which obtains a row echelon form, is called

Gaussian elimination or *row reduction*. For some problems, a row echelon form is sufficient, and it is a bit less work to only do this first part.

To attain the row echelon form we work systematically. We go column by column, starting at the first column. We find topmost entry in the first column that is not zero, and we call it the *pivot*. If there is no nonzero entry we move to the next column. We swap rows to put the row with the pivot as the first row. We divide the first row by the pivot to make the pivot entry be a 1. Now look at all the rows below and subtract the correct multiple of the pivot row so that all the entries below the pivot become zero.

After this procedure we forget that we had a first row (it is now fixed), and we forget about the column with the pivot and all the preceding zero columns. Below the pivot row, all the entries in these columns are just zero. Then we focus on the smaller matrix and we repeat the steps above.

It is best shown by example, so let us go back to the example from the beginning of the section. We keep the vertical line in the matrix, even though the procedure works on any matrix, not just an augmented matrix. We start with the first column and we locate the pivot, in this case the first entry of the first column.

$$\left[\begin{array}{ccc|c} 2 & 2 & 2 & 2 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{array} \right]$$

We multiply the first row by $1/2$.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{array} \right]$$

We subtract the first row from the second and third row (two elementary operations).

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 4 \\ 0 & 3 & 0 & 9 \end{array} \right]$$

We are done with the first column and the first row for now. We almost pretend the matrix doesn't have the first column and the first row.

$$\left[\begin{array}{ccc|c} * & * & * & * \\ * & 0 & 2 & 4 \\ * & 3 & 0 & 9 \end{array} \right]$$

OK, look at the second column, and notice that now the pivot is in the third row.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 4 \\ 0 & 3 & 0 & 9 \end{array} \right]$$

We swap rows.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 3 & 0 & 9 \\ 0 & 0 & 2 & 4 \end{array} \right]$$

And we divide the pivot row by 3.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 2 & 4 \end{array} \right]$$

We do not need to subtract anything as everything below the pivot is already zero. We move on, we again start ignoring the second row and second column and focus on

$$\left[\begin{array}{ccc|c} * & * & * & * \\ * & * & * & * \\ * & * & 2 & 4 \end{array} \right].$$

We find the pivot, then divide that row by 2:

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 2 & 4 \end{array} \right] \rightarrow \left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{array} \right].$$

The matrix is now in row echelon form.

The equation corresponding to the last row is $x_3 = 2$. We know x_3 and we could substitute it into the first two equations to get equations for x_1 and x_2 . Then we could do the same thing with x_2 , until we solve for all 3 variables. This procedure is called *backsubstitution* and we can achieve it via elementary operations. We start from the lowest pivot (leading entry in the row echelon form) and subtract the right multiple from the row above to make all the entries above this pivot zero. Then we move to the next pivot and so on. After we are done, we will have a matrix in reduced row echelon form.

We continue our example. Subtract the last row from the first to get

$$\left[\begin{array}{ccc|c} 1 & 1 & 0 & -1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{array} \right].$$

The entry above the pivot in the second row is already zero. So we move onto the next pivot, the one in the second row. We subtract this row from the top row to get

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & -4 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{array} \right].$$

The matrix is in reduced row echelon form.

If we now write down the equations for x_1, x_2, x_3 , we find

$$x_1 = -4, \quad x_2 = 3, \quad x_3 = 2.$$

In other words, we have solved the system.

3.3.3 Non-unique solutions and inconsistent systems

It is possible that the solution of a linear system of equations is not unique, or that no solution exists. Suppose for a moment that the row echelon form we found was

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{array} \right].$$

Then we have an equation $0 = 1$ coming from the last row. That is impossible and the equations are *inconsistent*. There is no solution to $A\vec{x} = \vec{b}$.

On the other hand, if we find a row echelon form

$$\left[\begin{array}{ccc|c} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{array} \right],$$

then there is no issue with finding solutions. In fact, we will find way too many. Let us continue with backsubstitution (subtracting 3 times the third row from the first) to find the reduced row echelon form and let's mark the pivots.

$$\left[\begin{array}{ccc|c} \boxed{1} & 2 & 0 & -5 \\ 0 & 0 & \boxed{1} & 3 \\ 0 & 0 & 0 & 0 \end{array} \right]$$

The last row is all zeros; it just says $0 = 0$ and we ignore it. The two remaining equations are

$$x_1 + 2x_2 = -5, \quad x_3 = 3.$$

Let us solve for the variables that corresponded to the pivots, that is x_1 and x_3 as there was a pivot in the first column and in the third column:

$$\begin{aligned} x_1 &= -2x_2 - 5, \\ x_3 &= 3. \end{aligned}$$

The variable x_2 can be anything you wish and we still get a solution. The x_2 is called a *free variable*. There are infinitely many solutions, one for every choice of x_2 . For example, if we pick $x_2 = 0$, then $x_1 = -5$, and $x_3 = 3$ give a solution. But we also get a solution by picking say $x_2 = 1$, in which case $x_1 = -9$ and $x_3 = 3$, or by picking $x_2 = -5$ in which case $x_1 = 5$ and $x_3 = 3$.

The general idea is that if any row has all zeros in the columns corresponding to the variables, but a nonzero entry in the column corresponding to the right-hand side \vec{b} , then the system is inconsistent and has no solutions. In other words, the system is inconsistent if you find a pivot on the right side of the vertical line drawn in the augmented matrix. Otherwise, the system is consistent, and at least one solution exists.

If the system is consistent:

- (i) If every column corresponding to a variable has a pivot element, then the solution is unique.
- (ii) If there are columns corresponding to variables with no pivot, then those are *free variables* that can be chosen arbitrarily, and there are infinitely many solutions.

When $\vec{b} = \vec{0}$, we have a so-called *homogeneous matrix equation*

$$A\vec{x} = \vec{0}.$$

There is no need to write an augmented matrix in this case. As the elementary operations do not do anything to a zero column, it always stays a zero column. Moreover, $A\vec{x} = \vec{0}$ always has at least one solution, namely $\vec{x} = \vec{0}$. Such a system is always consistent. It may have other solutions: If you find any free variables, then you get infinitely many solutions.

The set of solutions of $A\vec{x} = \vec{0}$ comes up quite often so people give it a name. It is called the *nullspace* or the *kernel* of A . One place where the kernel comes up is invertibility of a square matrix A . If the kernel of A contains a nonzero vector, then it contains infinitely many vectors (there was a free variable). But then it is impossible to invert $\vec{0}$, since infinitely many vectors go to $\vec{0}$, so there is no unique vector that A takes to $\vec{0}$. So if the kernel is nontrivial, that is, if there are any nonzero vectors, in other words, if there are any free variables, or in yet other words, if the row echelon form of A has columns without pivots, then A is not invertible. We will return to this idea later.

3.3.4 Computing the inverse

If the matrix A is square and there exists a unique solution \vec{x} to $A\vec{x} = \vec{b}$ for any \vec{b} (there are no free variables), then A is invertible.

In particular, if $A\vec{x} = \vec{b}$ then $\vec{x} = A^{-1}\vec{b}$. Now we just need to compute what A^{-1} is. We can surely do elimination every time we want to find $A^{-1}\vec{b}$, but that would be ridiculous. The mapping A^{-1} is linear and hence given by a matrix, and we have seen that to figure out the matrix we just need to find where does A^{-1} take the standard basis vectors $\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n$.

That is, to find the first column of A^{-1} we solve $A\vec{x} = \vec{e}_1$, because then $A^{-1}\vec{e}_1 = \vec{x}$. To find the second column of A^{-1} we solve $A\vec{x} = \vec{e}_2$. And so on. It is really just n eliminations that we need to do. But it gets even easier. If you think about it, the elimination is the same for everything on the left side of the augmented matrix. Doing n eliminations separately we would redo most of the computations. Best is to do all at once.

Therefore, to find the inverse of A , we write an $n \times 2n$ augmented matrix $[A | I]$, where I is the identity matrix, whose columns are precisely the standard basis vectors. We then perform row reduction until we arrive at the reduced row echelon form. If A is invertible, then pivots can be found in every column of A , and so the reduced row echelon form of $[A | I]$ looks like $[I | A^{-1}]$. We then just read off the inverse A^{-1} . If you do not find a pivot in every one of the first n columns of the augmented matrix, then A is not invertible.

This is best seen by example.

Example 3.3.2: Find the inverse of the matrix

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix}.$$

Solution: We write the augmented matrix and we start reducing:

$$\begin{array}{l} \left[\begin{array}{ccc|cccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 & 1 & 0 \\ 3 & 1 & 0 & 0 & 0 & 1 \end{array} \right] \rightarrow \left[\begin{array}{ccc|cccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -4 & -5 & -2 & 1 & 0 \\ 0 & -5 & -9 & -3 & 0 & 1 \end{array} \right] \rightarrow \left[\begin{array}{ccc|cccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 5/4 & 1/2 & 1/4 & 0 \\ 0 & -5 & -9 & -3 & 0 & 1 \end{array} \right] \rightarrow \\ \rightarrow \left[\begin{array}{ccc|cccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 5/4 & 1/2 & 1/4 & 0 \\ 0 & 0 & -11/4 & -1/2 & -5/4 & 1 \end{array} \right] \rightarrow \left[\begin{array}{ccc|cccc} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 5/4 & 1/2 & 1/4 & 0 \\ 0 & 0 & 1 & 2/11 & 5/11 & -4/11 \end{array} \right] \rightarrow \\ \rightarrow \left[\begin{array}{ccc|cccc} 1 & 2 & 0 & 5/11 & -5/11 & 12/11 \\ 0 & 1 & 0 & 3/11 & -9/11 & 5/11 \\ 0 & 0 & 1 & 2/11 & 5/11 & -4/11 \end{array} \right] \rightarrow \left[\begin{array}{ccc|cccc} 1 & 0 & 0 & -1/11 & 3/11 & 2/11 \\ 0 & 1 & 0 & 3/11 & -9/11 & 5/11 \\ 0 & 0 & 1 & 2/11 & 5/11 & -4/11 \end{array} \right]. \end{array}$$

So

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix}^{-1} = \begin{bmatrix} -1/11 & 3/11 & 2/11 \\ 3/11 & -9/11 & 5/11 \\ 2/11 & 5/11 & -4/11 \end{bmatrix}.$$

□

Not too terrible, no? Perhaps harder than inverting a 2×2 matrix for which we had a formula, but not too bad. Really in practice this is done efficiently by a computer.

3.3.5 Exercises

Exercise 3.3.1: Compute the reduced row echelon form for the following matrices:

a) $\begin{bmatrix} 1 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ b) $\begin{bmatrix} 3 & 3 \\ 6 & -3 \end{bmatrix}$ c) $\begin{bmatrix} 3 & 6 \\ -2 & -3 \end{bmatrix}$ d) $\begin{bmatrix} 6 & 6 & 7 & 7 \\ 1 & 1 & 0 & 1 \end{bmatrix}$

e) $\begin{bmatrix} 9 & 3 & 0 & 2 \\ 8 & 6 & 3 & 6 \\ 7 & 9 & 7 & 9 \end{bmatrix}$ f) $\begin{bmatrix} 2 & 1 & 3 & -3 \\ 6 & 0 & 0 & -1 \\ -2 & 4 & 4 & 3 \end{bmatrix}$ g) $\begin{bmatrix} 6 & 6 & 5 \\ 0 & -2 & 2 \\ 6 & 5 & 6 \end{bmatrix}$ h) $\begin{bmatrix} 0 & 2 & 0 & -1 \\ 6 & 6 & -3 & 3 \\ 6 & 2 & -3 & 5 \end{bmatrix}$

Exercise 3.3.2:* Compute the reduced row echelon form for the following matrices:

a) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ b) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ c) $\begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix}$ d) $\begin{bmatrix} 1 & -3 & 1 \\ 4 & 6 & -2 \\ -2 & 6 & -2 \end{bmatrix}$

e) $\begin{bmatrix} 2 & 2 & 5 & 2 \\ 1 & -2 & 4 & -1 \\ 0 & 3 & 1 & -2 \end{bmatrix}$ f) $\begin{bmatrix} -2 & 6 & 4 & 3 \\ 6 & 0 & -3 & 0 \\ 4 & 2 & -1 & 1 \end{bmatrix}$ g) $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ h) $\begin{bmatrix} 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 5 \end{bmatrix}$

Exercise 3.3.3: Compute the inverse of the given matrices

$$a) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$b) \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$c) \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix}$$

Exercise 3.3.4:* Compute the inverse of the given matrices

$$a) \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$b) \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$c) \begin{bmatrix} 2 & 4 & 0 \\ 2 & 2 & 3 \\ 2 & 4 & 1 \end{bmatrix}$$

Exercise 3.3.5: Solve (find all solutions), or show no solution exists

$$a) \begin{array}{l} 4x_1 + 3x_2 = -2 \\ -x_1 + x_2 = 4 \end{array}$$

$$\begin{array}{l} x_1 + 5x_2 + 3x_3 = 7 \\ 8x_1 + 7x_2 + 8x_3 = 8 \\ 4x_1 + 8x_2 + 6x_3 = 4 \end{array}$$

$$c) \begin{array}{l} 4x_1 + 8x_2 + 2x_3 = 3 \\ -x_1 - 2x_2 + 3x_3 = 1 \\ 4x_1 + 8x_2 = 2 \end{array}$$

$$\begin{array}{l} x + 2y + 3z = 4 \\ 2x - y + 3z = 1 \\ 3x + y + 6z = 6 \end{array}$$

Exercise 3.3.6:* Solve (find all solutions), or show no solution exists

$$a) \begin{array}{l} 4x_1 + 3x_2 = -1 \\ 5x_1 + 6x_2 = 4 \end{array}$$

$$\begin{array}{l} 5x + 6y + 5z = 7 \\ 6x + 8y + 6z = -1 \\ 5x + 2y + 5z = 2 \end{array}$$

$$c) \begin{array}{l} a + b + c = -1 \\ a + 5b + 6c = -1 \\ -2a + 5b + 6c = 8 \end{array}$$

$$\begin{array}{l} -2x_1 + 2x_2 + 8x_3 = 6 \\ x_2 + x_3 = 2 \\ x_1 + 4x_2 + x_3 = 7 \end{array}$$

Exercise 3.3.7: By computing the inverse, solve the following systems for \vec{x} .

$$a) \begin{bmatrix} 4 & 1 \\ -1 & 3 \end{bmatrix} \vec{x} = \begin{bmatrix} 13 \\ 26 \end{bmatrix}$$

$$b) \begin{bmatrix} 3 & 3 \\ 3 & 4 \end{bmatrix} \vec{x} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

Exercise 3.3.8:* By computing the inverse, solve the following systems for \vec{x} .

$$a) \begin{bmatrix} -1 & 1 \\ 3 & 3 \end{bmatrix} \vec{x} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$

$$b) \begin{bmatrix} 2 & 7 \\ 1 & 6 \end{bmatrix} \vec{x} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

3.4 Subspaces and dimension

Attribution: [JL], §A.4.

Learning Objectives

After this section, you will be able to:

- Determine if a set of vectors is linearly independent,
- Compute the rank of a matrix,
- Find a maximal linearly independent subset of a set of vectors, and
- Compute a basis of a subspace and the dimension of that subspace.

3.4.1 Linear independence and rank

If rows of a matrix correspond to equations, it might be good to find out how many equations do we really need to find the same set of solutions. Similarly, if we find a number of solutions to a linear equation $A\vec{x} = \vec{0}$, we may ask if we found enough so that all other solutions can be formed out of the given set. The concept we want is that of linear independence. The same concept is useful for differential equations, for example in [chapter 2](#).

Definition 3.4.1

Given row or column vectors $\vec{y}_1, \vec{y}_2, \dots, \vec{y}_n$, a *linear combination* is an expression of the form

$$\alpha_1\vec{y}_1 + \alpha_2\vec{y}_2 + \cdots + \alpha_n\vec{y}_n,$$

where $\alpha_1, \alpha_2, \dots, \alpha_n$ are all scalars.

For example, $3\vec{y}_1 + \vec{y}_2 - 5\vec{y}_3$ is a linear combination of \vec{y}_1, \vec{y}_2 , and \vec{y}_3 .

We have seen linear combinations before. The expression

$$A\vec{x}$$

is a linear combination of the columns of A , while

$$\vec{x}^T A = (A^T \vec{x})^T$$

is a linear combination of the rows of A .

The way linear combinations come up in our study of differential equations is similar to the following computation. Suppose that $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$ are solutions to $A\vec{x}_1 = \vec{0}, A\vec{x}_2 = \vec{0}, \dots, A\vec{x}_n = \vec{0}$. Then the linear combination

$$\vec{y} = \alpha_1\vec{x}_1 + \alpha_2\vec{x}_2 + \cdots + \alpha_n\vec{x}_n$$

is a solution to $A\vec{y} = \vec{0}$:

$$\begin{aligned} A\vec{y} &= A(\alpha_1\vec{x}_1 + \alpha_2\vec{x}_2 + \cdots + \alpha_n\vec{x}_n) = \\ &= \alpha_1 A\vec{x}_1 + \alpha_2 A\vec{x}_2 + \cdots + \alpha_n A\vec{x}_n = \alpha_1 \vec{0} + \alpha_2 \vec{0} + \cdots + \alpha_n \vec{0} = \vec{0}. \end{aligned}$$

So if you have found enough solutions, you have them all. The question is, when did we find enough of them?

Definition 3.4.2

We say the vectors $\vec{y}_1, \vec{y}_2, \dots, \vec{y}_n$ are *linearly independent* if the only solution to

$$\alpha_1 \vec{x}_1 + \alpha_2 \vec{x}_2 + \cdots + \alpha_n \vec{x}_n = \vec{0}$$

is $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$. Otherwise, we say the vectors are *linearly dependent*.

Example 3.4.1: The vectors $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ are linearly independent.

Solution: Let's try:

$$\alpha_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \alpha_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ 2\alpha_1 + \alpha_2 \end{bmatrix} = \vec{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

So $\alpha_1 = 0$, and then it is clear that $\alpha_2 = 0$ as well. In other words, the vectors are linearly independent. \square

If a set of vectors is linearly dependent, that is, some of the α_j 's are nonzero, then we can solve for one vector in terms of the others. Suppose $\alpha_1 \neq 0$. Since $\alpha_1 \vec{x}_1 + \alpha_2 \vec{x}_2 + \cdots + \alpha_n \vec{x}_n = \vec{0}$, then

$$\vec{x}_1 = \frac{-\alpha_2}{\alpha_1} \vec{x}_2 - \frac{-\alpha_3}{\alpha_1} \vec{x}_3 + \cdots + \frac{-\alpha_n}{\alpha_1} \vec{x}_n.$$

For example,

$$2 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - 4 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + 2 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

and so

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}.$$

You may have noticed that solving for those α_j 's is just solving linear equations, and so you may not be surprised that to check if a set of vectors is linearly independent we use row reduction.

Given a set of vectors, we may not be interested in just finding if they are linearly independent or not, we may be interested in finding a linearly independent subset. Or perhaps we may want to find some other vectors that give the same linear combinations and are linearly independent. The way to figure this out is to form a matrix out of our vectors. If we have row vectors we consider them as rows of a matrix. If we have column vectors we consider them columns of a matrix.

Definition 3.4.3

The set of all linear combinations of a set of vectors is called their *span*.

$$\text{span}\{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\} = \{\text{Set of all linear combinations of } \vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\}.$$

Definition 3.4.4

Given a matrix A , the maximal number of linearly independent rows is called the *rank* of A , and we write “rank A ” for the rank.

For example,

$$\text{rank} \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & -1 \end{bmatrix} = 1.$$

The second and third row are multiples of the first one. We cannot choose more than one row and still have a linearly independent set. But what is

$$\text{rank} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = ?$$

That seems to be a tougher question to answer. The first two rows are linearly independent, so the rank is at least two. If we would set up the equations for the α_1 , α_2 , and α_3 , we would find a system with infinitely many solutions. One solution is

$$[1 \ 2 \ 3] - 2[4 \ 5 \ 6] + [7 \ 8 \ 9] = [0 \ 0 \ 0].$$

So the set of all three rows is linearly dependent, the rank cannot be 3. Therefore the rank is 2.

But how can we do this in a more systematic way? We find the row echelon form!

$$\text{Row echelon form of } \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \text{ is } \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{bmatrix}.$$

The elementary row operations do not change the set of linear combinations of the rows (that was one of the main reasons for defining them as they were). In other words, the span of the rows of the A is the same as the span of the rows of the row echelon form of A . In particular, the number of linearly independent rows is the same. And in the row echelon form, all nonzero rows are linearly independent. This is not hard to see. Consider the two nonzero rows in the example above. Suppose we tried to solve for the α_1 and α_2 in

$$\alpha_1 [1 \ 2 \ 3] + \alpha_2 [0 \ 1 \ 2] = [0 \ 0 \ 0].$$

Since the first column of the row echelon matrix has zeros except in the first row means that $\alpha_1 = 0$. For the same reason, α_2 is zero. We only have two nonzero rows, and they are linearly independent, so the rank of the matrix is 2.

The span of the rows is called the *row space*. The row space of A and the row echelon form of A are the same. In the example,

$$\begin{aligned} \text{row space of } \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} &= \text{span} \{ [1 \ 2 \ 3], [4 \ 5 \ 6], [7 \ 8 \ 9] \} \\ &= \text{span} \{ [1 \ 2 \ 3], [0 \ 1 \ 2] \}. \end{aligned}$$

Similarly to row space, the span of columns is called the *column space*.

$$\text{column space of } \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \text{span} \left\{ \begin{bmatrix} 1 \\ 4 \\ 7 \end{bmatrix}, \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix}, \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix} \right\}.$$

So it may also be good to find the number of linearly independent columns of A . One way to do that is to find the number of linearly independent rows of A^T . It is a tremendously useful fact that the number of linearly independent columns is always the same as the number of linearly independent rows:

Theorem 3.4.1

$$\text{rank } A = \text{rank } A^T$$

In particular, to find a set of linearly independent columns we need to look at where the pivots were. If you recall above, when solving $A\vec{x} = \vec{0}$ the key was finding the pivots, any non-pivot columns corresponded to free variables. That means we can solve for the non-pivot columns in terms of the pivot columns. Let's see an example.

Example 3.4.2: Find the linearly independent columns of the matrix

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix}.$$

Solution: We find a pivot and reduce the rows below:

$$\begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix} \rightarrow \begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 0 & 0 & -1 & -2 \\ 3 & 6 & 7 & 8 \end{bmatrix} \rightarrow \begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 0 & 0 & -1 & -2 \\ 0 & 0 & -2 & -4 \end{bmatrix}.$$

We find the next pivot, make it one, and rinse and repeat:

$$\begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 0 & 0 & \boxed{-1} & -2 \\ 0 & 0 & -2 & -4 \end{bmatrix} \rightarrow \begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 0 & 0 & \boxed{1} & 2 \\ 0 & 0 & -2 & -4 \end{bmatrix} \rightarrow \begin{bmatrix} \boxed{1} & 2 & 3 & 4 \\ 0 & 0 & \boxed{1} & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The final matrix is the row echelon form of the matrix. Consider the pivots that we marked. The pivot columns are the first and the third column. All other columns correspond to free variables when solving $A\vec{x} = \vec{0}$, so all other columns can be solved in terms of the first and the third column. In other words

$$\text{column space of } \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix} = \text{span} \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}, \begin{bmatrix} 4 \\ 6 \\ 8 \end{bmatrix} \right\} = \text{span} \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix} \right\}.$$

We could perhaps use another pair of columns to get the same span, but the first and the third are guaranteed to work because they are pivot columns.

The discussion above could be expanded into a proof of the theorem if we wanted. As each nonzero row in the row echelon form contains a pivot, then the rank is the number of pivots, which is the same as the maximal number of linearly independent columns.

The idea also works in reverse. Suppose we have a bunch of column vectors and we just need to find a linearly independent set. For example, suppose we started with the vectors

$$\vec{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \vec{v}_2 = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}, \quad \vec{v}_3 = \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}, \quad \vec{v}_4 = \begin{bmatrix} 4 \\ 6 \\ 8 \end{bmatrix}.$$

These vectors are not linearly independent as we saw above. In particular, the span \vec{v}_1 and \vec{v}_3 is the same as the span of all four of the vectors. So \vec{v}_2 and \vec{v}_4 can both be written as linear combinations of \vec{v}_1 and \vec{v}_3 . A common thing that comes up in practice is that one gets a set of vectors whose span is the set of solutions of some problem. But perhaps we get way too many vectors, we want to simplify. For example above, all vectors in the span of $\vec{v}_1, \vec{v}_2, \vec{v}_3, \vec{v}_4$ can be written $\alpha_1\vec{v}_1 + \alpha_2\vec{v}_2 + \alpha_3\vec{v}_3 + \alpha_4\vec{v}_4$ for some numbers $\alpha_1, \alpha_2, \alpha_3, \alpha_4$. But it is also true that every such vector can be written as $a\vec{v}_1 + b\vec{v}_3$ for two numbers a and b . And one has to admit, that looks much simpler. Moreover, these numbers a and b are unique. More on that later in this section.

To find this linearly independent set we simply take our vectors and form the matrix $[\vec{v}_1 \ \vec{v}_2 \ \vec{v}_3 \ \vec{v}_4]$, that is, the matrix

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix}.$$

We crank up the row-reduction machine, feed this matrix into it, and find the pivot columns and pick those. In this case, \vec{v}_1 and \vec{v}_3 .

3.4.2 Subspaces, basis, and dimension

We often find ourselves looking at the set of solutions of a linear equation $L\vec{x} = \vec{0}$ for some matrix L , that is, we are interested in the kernel of L . The set of all such solutions has a nice structure: It looks and acts a lot like some euclidean space \mathbb{R}^k .

We say that a set S of vectors in \mathbb{R}^n is a *subspace* if whenever \vec{x} and \vec{y} are members of S and α is a scalar, then

$$\vec{x} + \vec{y}, \quad \text{and} \quad \alpha\vec{x}$$

are also members of S . That is, we can add and multiply by scalars and we still land in S . So every linear combination of vectors of S is still in S . That is really what a subspace is. It is a subset where we can take linear combinations and still end up being in the subset. Consequently the span of a number of vectors is automatically a subspace.

Example 3.4.3: If we let $S = \mathbb{R}^n$, then this S is a subspace of \mathbb{R}^n . Adding any two vectors in \mathbb{R}^n gets a vector in \mathbb{R}^n , and so does multiplying by scalars.

The set $S' = \{\vec{0}\}$, that is, the set of the zero vector by itself, is also a subspace of \mathbb{R}^n . There is only one vector in this subspace, so we only need to check for that one vector, and everything checks out: $\vec{0} + \vec{0} = \vec{0}$ and $\alpha\vec{0} = \vec{0}$.

The set S'' of all the vectors of the form (a, a) for any real number a , such as $(1, 1)$, $(3, 3)$, or $(-0.5, -0.5)$ is a subspace of \mathbb{R}^2 . Adding two such vectors, say $(1, 1) + (3, 3) = (4, 4)$ again gets a vector of the same form, and so does multiplying by a scalar, say $8(1, 1) = (8, 8)$.

If S is a subspace and we can find k linearly independent vectors in S

$$\vec{v}_1, \vec{v}_2, \dots, \vec{v}_k,$$

such that every other vector in S is a linear combination of $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_k$, then the set $\{\vec{v}_1, \vec{v}_2, \dots, \vec{v}_k\}$ is called a *basis* of S . In other words, S is the span of $\{\vec{v}_1, \vec{v}_2, \dots, \vec{v}_k\}$. We say that S has dimension k , and we write

$$\dim S = k.$$

Theorem 3.4.2

If $S \subset \mathbb{R}^n$ is a subspace and S is not the trivial subspace $\{\vec{0}\}$, then there exists a unique positive integer k (the dimension) and a (not unique) basis $\{\vec{v}_1, \vec{v}_2, \dots, \vec{v}_k\}$, such that every \vec{w} in S can be uniquely represented by

$$\vec{w} = \alpha_1 \vec{v}_1 + \alpha_2 \vec{v}_2 + \cdots + \alpha_k \vec{v}_k,$$

for some scalars $\alpha_1, \alpha_2, \dots, \alpha_k$.

Just like a vector in \mathbb{R}^k is represented by a k -tuple of numbers, so is a vector in a k -dimensional subspace of \mathbb{R}^n represented by a k -tuple of numbers. At least once we have fixed a basis. A different basis would give a different k -tuple of numbers for the same vector.

We should reiterate that while k is unique (a subspace cannot have two different dimensions), the set of basis vectors is not at all unique. There are lots of different bases for any given subspace. Finding just the right basis for a subspace is a large part of what one does in linear algebra. In fact, that is what we spend a lot of time on in linear differential equations, although at first glance it may not seem like that is what we are doing.

Example 3.4.4: The standard basis

$$\vec{e}_1, \vec{e}_2, \dots, \vec{e}_n,$$

is a basis of \mathbb{R}^n , (hence the name). So as expected

$$\dim \mathbb{R}^n = n.$$

On the other hand the subspace $\{\vec{0}\}$ is of dimension 0.

The subspace S'' from a previous example, that is, the set of vectors (a, a) is of dimension 1. One possible basis is simply $\{(1, 1)\}$, the single vector $(1, 1)$: every vector in S'' can be represented by $a(1, 1) = (a, a)$. Similarly another possible basis would be $\{(-1, -1)\}$. Then the vector (a, a) would be represented as $(-a)(1, 1)$.

Row and column spaces of a matrix are also examples of subspaces, as they are given as the span of vectors. We can use what we know about rank, row spaces, and column spaces from the previous section to find a basis.

Example 3.4.5: In the last section, we considered the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix}.$$

Using row reduction to find the pivot columns, we found

$$\text{column space of } A \left(\begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 3 & 6 & 7 & 8 \end{bmatrix} \right) = \text{span} \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix} \right\}.$$

What we did was we found the basis of the column space. The basis has two elements, and so the column space of A is two dimensional. Notice that the rank of A is two.

We would have followed the same procedure if we wanted to find the basis of the subspace X spanned by

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}, \begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}, \begin{bmatrix} 4 \\ 6 \\ 8 \end{bmatrix}.$$

We would have simply formed the matrix A with these vectors as columns and repeated the computation above. The subspace X is then the column space of A .

Example 3.4.6: Consider the matrix

$$L = \begin{bmatrix} 1 & 2 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 5 \end{bmatrix}$$

Conveniently, the matrix is in reduced row echelon form. The matrix is of rank 3. The column space is the span of the pivot columns. It is the 3-dimensional space

$$\text{column space of } L = \text{span} \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\} = \mathbb{R}^3.$$

The row space is the 3-dimensional space

$$\text{row space of } L = \text{span} \{ [1 \ 2 \ 0 \ 0 \ 3], [0 \ 0 \ 1 \ 0 \ 4], [0 \ 0 \ 0 \ 1 \ 5] \}.$$

As these vectors have 5 components, we think of the row space of L as a subspace of \mathbb{R}^5 .

The way the dimensions worked out in the examples is not an accident. Since the number of vectors that we needed to take was always the same as the number of pivots, and the number of pivots is the rank, we get the following result.

Theorem 3.4.3 (Rank)

The dimension of the column space and the dimension of the row space of a matrix A are both equal to the rank of A .

3.4.3 Exercises

Exercise 3.4.1: Compute the rank of the given matrices

$$a) \begin{bmatrix} 6 & 3 & 5 \\ 1 & 4 & 1 \\ 7 & 7 & 6 \end{bmatrix}$$

$$b) \begin{bmatrix} 5 & -2 & -1 \\ 3 & 0 & 6 \\ 2 & 4 & 5 \end{bmatrix}$$

$$c) \begin{bmatrix} 1 & 2 & 3 \\ -1 & -2 & -3 \\ 2 & 4 & 6 \end{bmatrix}$$

Exercise 3.4.2:* Compute the rank of the given matrices

$$a) \begin{bmatrix} 7 & -1 & 6 \\ 7 & 7 & 7 \\ 7 & 6 & 2 \end{bmatrix}$$

$$b) \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 2 & 2 & 2 \end{bmatrix}$$

$$c) \begin{bmatrix} 0 & 3 & -1 \\ 6 & 3 & 1 \\ 4 & 7 & -1 \end{bmatrix}$$

Exercise 3.4.3: For the matrices in [Exercise 3.4.1](#), find a linearly independent set of row vectors that span the row space (they don't need to be rows of the matrix).

Exercise 3.4.4: For the matrices in [Exercise 3.4.1](#), find a linearly independent set of columns that span the column space. That is, find the pivot columns of the matrices.

Exercise 3.4.5:* For the matrices in [Exercise 3.4.2](#), find a linearly independent set of row vectors that span the row space (they don't need to be rows of the matrix).

Exercise 3.4.6:* For the matrices in [Exercise 3.4.2](#), find a linearly independent set of columns that span the column space. That is, find the pivot columns of the matrices.

Exercise 3.4.7: Find a linearly independent subset of the following vectors that has the same span.

$$\begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ -2 \\ -4 \end{bmatrix}, \quad \begin{bmatrix} -2 \\ 4 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} -1 \\ 3 \\ -2 \end{bmatrix}$$

Exercise 3.4.8:* Find a linearly independent subset of the following vectors that has the same span.

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 1 \\ -5 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 3 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} -3 \\ 2 \\ 4 \end{bmatrix}$$

Exercise 3.4.9: For the following sets of vectors, determine if the set is linearly independent. Then find a basis for the subspace spanned by the vectors, and find the dimension of the subspace.

$$a) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix}$$

$$b) \begin{bmatrix} 1 \\ 0 \\ 5 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix}$$

$$c) \begin{bmatrix} -4 \\ -3 \\ 5 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 3 \\ 3 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix}$$

$$d) \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} -1 \\ -1 \\ 2 \end{bmatrix}$$

$$e) \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 2 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix}$$

$$f) \begin{bmatrix} 3 \\ 1 \\ 3 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 4 \\ -4 \end{bmatrix}, \quad \begin{bmatrix} -5 \\ -5 \\ -2 \end{bmatrix}$$

Exercise 3.4.10:* For the following sets of vectors, determine if the set is linearly independent. Then find a basis for the subspace spanned by the vectors, and find the dimension of the subspace.

$$a) \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$b) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

$$c) \begin{bmatrix} 5 \\ 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 5 \\ -1 \\ 5 \end{bmatrix}, \begin{bmatrix} -1 \\ 3 \\ -4 \end{bmatrix}$$

$$d) \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ 4 \\ -3 \end{bmatrix}$$

$$e) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}$$

$$f) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

Exercise 3.4.11: Suppose that X is the set of all the vectors of \mathbb{R}^3 whose third component is zero. Is X a subspace? And if so, find a basis and the dimension.

3.5 Determinant

Attribution: [JL], §A.6.

Learning Objectives

After this section, you will be able to:

- Compute the determinant of a 2×2 matrix,
- Use cofactor expansion to compute the determinant of larger matrices, and
- Use the determinant to make statements about invertibility or rank of a matrix, and linear independence of the columns of that matrix.

For square matrices we define a useful quantity called the *determinant*. We define the determinant of a 1×1 matrix as the value of its only entry

$$\det([a]) \stackrel{\text{def}}{=} a.$$

For a 2×2 matrix we define

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} \stackrel{\text{def}}{=} ad - bc.$$

Before defining the determinant for larger matrices, we note the meaning of the determinant. An $n \times n$ matrix gives a mapping of the n -dimensional Euclidean space \mathbb{R}^n to itself. In particular, a 2×2 matrix A is a mapping of the plane to itself. The determinant of A is the factor by which the area of objects changes. If we take the unit square (square of side 1) in the plane, then A takes the square to a parallelogram of area $|\det(A)|$. The sign of $\det(A)$ denotes a change of orientation (negative if the axes get flipped). For example, let

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

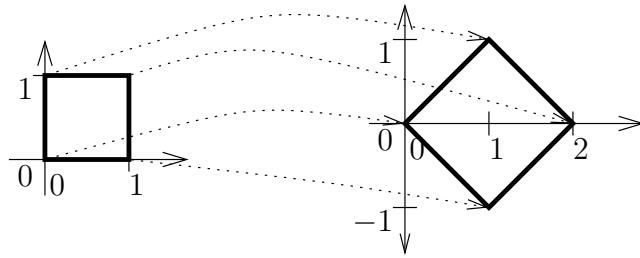
Then $\det(A) = 1 + 1 = 2$. Let us see where A sends the unit square with vertices $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$. The point $(0, 0)$ gets sent to $(0, 0)$.

$$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

The image of the square is another square with vertices $(0, 0)$, $(1, -1)$, $(1, 1)$, and $(2, 0)$. The image square has a side of length $\sqrt{2}$ and is therefore of area 2. See Figure 3.5 on the facing page.

In general the image of a square is going to be a parallelogram. In high school geometry, you may have seen a formula for computing the area of a parallelogram with vertices $(0, 0)$, (a, c) , (b, d) and $(a + b, c + d)$. The area is

$$\left| \det\begin{pmatrix} a & b \\ c & d \end{pmatrix} \right| = |ad - bc|.$$

Figure 3.5: Image of the unit square via the mapping A .

The vertical lines above mean absolute value. The matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ carries the unit square to the given parallelogram.

There are a number of ways to define the determinant for an $n \times n$ matrix. Let us use the so-called *cofactor expansion*. We define A_{ij} as the matrix A with the i^{th} row and the j^{th} column deleted. For example, if

$$\text{If } A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \quad \text{then } A_{12} = \begin{bmatrix} 4 & 6 \\ 7 & 9 \end{bmatrix} \quad \text{and } A_{23} = \begin{bmatrix} 1 & 2 \\ 7 & 8 \end{bmatrix}.$$

We now define the determinant recursively

$$\det(A) \stackrel{\text{def}}{=} \sum_{j=1}^n (-1)^{1+j} a_{1j} \det(A_{1j}),$$

or in other words

$$\det(A) = a_{11} \det(A_{11}) - a_{12} \det(A_{12}) + a_{13} \det(A_{13}) - \dots \begin{cases} +a_{1n} \det(A_{1n}) & \text{if } n \text{ is odd,} \\ -a_{1n} \det(A_{1n}) & \text{if } n \text{ even.} \end{cases}$$

For a 3×3 matrix, we get $\det(A) = a_{11} \det(A_{11}) - a_{12} \det(A_{12}) + a_{13} \det(A_{13})$. For example,

$$\begin{aligned} \det \left(\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \right) &= 1 \cdot \det \left(\begin{bmatrix} 5 & 6 \\ 8 & 9 \end{bmatrix} \right) - 2 \cdot \det \left(\begin{bmatrix} 4 & 6 \\ 7 & 9 \end{bmatrix} \right) + 3 \cdot \det \left(\begin{bmatrix} 4 & 5 \\ 7 & 8 \end{bmatrix} \right) \\ &= 1(5 \cdot 9 - 6 \cdot 8) - 2(4 \cdot 9 - 6 \cdot 7) + 3(4 \cdot 8 - 5 \cdot 7) = 0. \end{aligned}$$

It turns out that we did not have to necessarily use the first row. That is for any i ,

$$\det(A) = \sum_{j=1}^n (-1)^{i+j} a_{ij} \det(A_{ij}).$$

It is sometimes useful to use a row other than the first. In the following example it is more convenient to expand along the second row. Notice that for the second row we are starting

with a negative sign.

$$\begin{aligned}\det\left(\begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 0 \\ 7 & 8 & 9 \end{bmatrix}\right) &= -0 \cdot \det\left(\begin{bmatrix} 2 & 3 \\ 8 & 9 \end{bmatrix}\right) + 5 \cdot \det\left(\begin{bmatrix} 1 & 3 \\ 7 & 9 \end{bmatrix}\right) - 0 \cdot \det\left(\begin{bmatrix} 1 & 2 \\ 7 & 8 \end{bmatrix}\right) \\ &= 0 + 5(1 \cdot 9 - 3 \cdot 7) + 0 = -60.\end{aligned}$$

Let us check if it is really the same as expanding along the first row,

$$\begin{aligned}\det\left(\begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 0 \\ 7 & 8 & 9 \end{bmatrix}\right) &= 1 \cdot \det\left(\begin{bmatrix} 5 & 0 \\ 8 & 9 \end{bmatrix}\right) - 2 \cdot \det\left(\begin{bmatrix} 0 & 0 \\ 7 & 9 \end{bmatrix}\right) + 3 \cdot \det\left(\begin{bmatrix} 0 & 5 \\ 7 & 8 \end{bmatrix}\right) \\ &= 1(5 \cdot 9 - 0 \cdot 8) - 2(0 \cdot 9 - 0 \cdot 7) + 3(0 \cdot 8 - 5 \cdot 7) = -60.\end{aligned}$$

In computing the determinant, we alternately add and subtract the determinants of the submatrices A_{ij} multiplied by a_{ij} for a fixed i and all j . The numbers $(-1)^{i+j} \det(A_{ij})$ are called *cofactors* of the matrix. And that is why this method of computing the determinant is called the *cofactor expansion*.

Similarly we do not need to expand along a row, we can expand along a column. For any j

$$\det(A) = \sum_{i=1}^n (-1)^{i+j} a_{ij} \det(A_{ij}).$$

A related fact is that

$$\det(A) = \det(A^T).$$

Recall that a matrix is *upper triangular* if all elements below the main diagonal are 0. For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 5 & 6 \\ 0 & 0 & 9 \end{bmatrix}$$

is upper triangular. Similarly a *lower triangular* matrix is one where everything above the diagonal is zero. For example,

$$\begin{bmatrix} 1 & 0 & 0 \\ 4 & 5 & 0 \\ 7 & 8 & 9 \end{bmatrix}.$$

The determinant for triangular matrices is very simple to compute. Consider the lower triangular matrix. If we expand along the first row, we find that the determinant is 1 times the determinant of the lower triangular matrix $\begin{bmatrix} 5 & 0 \\ 8 & 9 \end{bmatrix}$. So the determinant is just the product of the diagonal entries:

$$\det\left(\begin{bmatrix} 1 & 0 & 0 \\ 4 & 5 & 0 \\ 7 & 8 & 9 \end{bmatrix}\right) = 1 \cdot 5 \cdot 9 = 45.$$

Similarly for upper triangular matrices

$$\det \begin{pmatrix} 1 & 2 & 3 \\ 0 & 5 & 6 \\ 0 & 0 & 9 \end{pmatrix} = 1 \cdot 5 \cdot 9 = 45.$$

In general, if A is triangular, then

$$\det(A) = a_{11}a_{22} \cdots a_{nn}.$$

If A is diagonal, then it is also triangular (upper and lower), so same formula applies. For example,

$$\det \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix} = 2 \cdot 3 \cdot 5 = 30.$$

In particular, the identity matrix I is diagonal, and the diagonal entries are all 1. Thus,

$$\det(I) = 1.$$

Another way that we can compute determinants is by using row reduction. Since the row echelon form is a diagonal matrix, this will make it easy to compute the determinant using the product of the diagonal entries. However, we need to know how the determinant is affected by elementary row operations.

Theorem 3.5.1 (Properties of the Determinant)

[detElem] Let A be a square $n \times n$ matrix.

1. If B obtained from A by interchanging two rows (or two columns) of A , then $\det(B) = -\det(A)$.
2. If B is obtained from A by multiplying a row of column by the number r , then $\det(B) = r \det(A)$.
3. If B is obtained from A by multiplying a row (or column) by a non-zero number r and adding the result to another row, then $\det(B) = \det(A)$.

Proof. The proof of each of these facts comes from the cofactor expansion of the determinant.

1. Assume that B is obtained by interchanging the first and second row of A . We will use cofactor expansion along the first row to find the determinant of A , and the second row for the determinant of B . We get that

$$\det(A) = \sum_{j=1}^n (-1)^{1+j} a_{1j} \det(A_{1j})$$

and

$$\det(B) = \sum_{j=1}^n (-1)^{2+j} b_{2j} \det(B_{2j}).$$

However, since the second row of B is the first row of A , we know that $b_{2j} = a_{1j}$ for all j . In addition, this swap means that we also have that $B_{2j} = A_{1j}$ for each of the cofactors in this expansion. All of these cofactor matrices are made up of the second through last rows of A , with the appropriate columns removed at each step.

Therefore, the only difference between these two formulas is that the A formula starts with $(-1)^{1+j}$ and the B formula starts with $(-1)^{2+j}$. Thus, $\det(B)$ will have an additional factor of -1 in it, giving the desired result.

The exact same process works for swapping any two adjacent rows of the matrix, giving that this also provides a -1 in the computation of the determinant. For non-adjacent rows, we use the fact that to any swap of non-adjacent rows of a matrix requires an *odd* number of adjacent row swaps. For example, if we want to swap rows 1 and 3, we can swap row 1 with row 2, then row 2 with row 3, and finally swap row 1 with row 2 again. This will put the first row in the third spot and the third row up in the first slot. Since each of these adjacent switches adds a minus sign, doing an odd number of switches still results in adding a single minus sign to the computation of the determinant.

- Assume that we want to multiply the k th row of A by the number r to get B . We use cofactor expansion along this same k th row to find the determinant of each matrix. We get that

$$\det(A) = \sum_{j=1}^n (-1)^{k+j} a_{kj} \det(A_{kj})$$

and

$$\det(B) = \sum_{j=1}^n (-1)^{k+j} b_{kj} \det(B_{kj}) = \sum_{j=1}^n (-1)^{k+j} r a_{kj} \det(B_{kj}).$$

However, the minor B_{kj} ignores the k th row of the matrix B , so the minors are identical to those of A . Thus, we have that

$$\det(B) = \sum_{j=1}^n (-1)^{k+j} r a_{kj} \det(B_{kj}) = r \sum_{j=1}^n (-1)^{k+j} a_{kj} \det(A_{kj}) = r \det(A).$$

- Assume that B is formed by adding r copies of the k th row of A to the i th row. Since the i th row is the one being changed, we will use cofactor expansion there to compute each determinant. We get that

$$\det(A) = \sum_{j=1}^n (-1)^{i+j} a_{ij} \det(A_{ij})$$

and

$$\det(B) = \sum_{j=1}^n (-1)^{i+j} b_{ij} \det(B_{ij}) = \sum_{j=1}^n (-1)^{i+j} (a_{ij} + r a_{kj}) \det(A_{ij})$$

where we have replaced the minors of B by the minors of A because they ignore the i th row, which is the only thing that has changed. We can now split the determinant

of B into two parts

$$\sum (-1)^{i+j}(a_{ij} + ra_{kj}) \det(A_{ij}) = \sum (-1)^{i+j}a_{ij} \det(A_{ij}) + \sum (-1)^{i+j}ra_{kj} \det(A_{ij}).$$

The first of these is the determinant of the matrix A . The second is the determinant of a new matrix that we will call C . C is the same as the matrix A , except that we have replaced the i th row of A by r times the k th row of A . Thus, the i th row of this matrix C is a multiple of the k th row. This means that the rows of C are not linearly independent. By [Theorem 3.5.4](#) coming up later (don't worry, it does not depend on this result), this tells us that the determinant of C is zero. Therefore

$$\det(B) = \det(A) + \det(C) = \det(A)$$

so this operation does not change the determinant of the matrix.

□

These correspond to the three elementary row operations that we use to row reduce matrices. In order to use this to compute determinants, we need to keep track of each of these operations and how the determinant changes at each step.

Example 3.5.1: Compute the determinant of the matrix

$$\begin{bmatrix} -4 & -2 & 4 \\ -3 & -3 & 2 \\ -2 & -3 & 1 \end{bmatrix}$$

using row reduction.

Solution: We will go through the process of row reduction to find the determinant. We need to keep track of each time that we swap rows (to add a minus sign) and that we multiply a row by a constant (to factor in that constant). Throughout this process, we will use A to refer to the initial matrix

$$A = \begin{bmatrix} -4 & -2 & 4 \\ -3 & -3 & 2 \\ -2 & -3 & 1 \end{bmatrix}$$

and M will refer to wherever we are in the process. So we will start by dividing the first row of the matrix by -4

$$\begin{bmatrix} -4 & -2 & 4 \\ -3 & -3 & 2 \\ -2 & -3 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1/2 & -1 \\ -3 & -3 & 2 \\ -2 & -3 & 1 \end{bmatrix}.$$

Since we divided by -4 , Theorem ?? tells us that

$$\det(M) = -\frac{1}{4} \det(A).$$

The next step of row reduction will be to use the 1 in the top left to cancel out the -3 and -2 below it. Part (c) in Theorem ?? says that this doesn't change the determinant. Therefore, the row reduction gives

$$\begin{bmatrix} 1 & 1/2 & -1 \\ -3 & -3 & 2 \\ -2 & -3 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1/2 & -1 \\ 0 & -3/2 & -1 \\ 0 & -2 & -1 \end{bmatrix}$$

and we still have that

$$\det(M) = -\frac{1}{4} \det(A).$$

Next, we will multiply row 2 by $-\frac{2}{3}$, which gives

$$\begin{bmatrix} 1 & 1/2 & -1 \\ 0 & -3/2 & -1 \\ 0 & -2 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1/2 & -1 \\ 0 & 1 & 2/3 \\ 0 & -2 & -1 \end{bmatrix}.$$

Adding this in to our previous steps using Theorem ??, we get that

$$\det(M) = \left(-\frac{2}{3}\right) \left(-\frac{1}{4}\right) \det(A).$$

Finally, we add two copies of row 2 to row 3, which does not change the determinant and gives the matrix

$$\begin{bmatrix} 1 & 1/2 & -1 \\ 0 & 1 & 2/3 \\ 0 & -2 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1/2 & -1 \\ 0 & 1 & 2/3 \\ 0 & 0 & 1/3 \end{bmatrix}$$

with

$$\det(M) = \left(-\frac{2}{3}\right) \left(-\frac{1}{4}\right) \det(A).$$

We can rearrange this expression to say that

$$\det(A) = 6 \det(M)$$

and we can easily compute that $\det(M) = \frac{1}{3}$ by multiplying the diagonal entries. Thus, we have that $\det(A) = 2$. □

Exercise 3.5.1: Compute $\det(A)$ using cofactor expansion and show that you get the same answer.

The determinant is telling you how geometric objects scale. If B doubles the sizes of geometric objects and A triples them, then AB (which applies B to an object and then it applies A) should make size go up by a factor of 6. This is true in general:

Theorem 3.5.2

$$\det(AB) = \det(A) \det(B).$$

This property is one of the most useful, and it is employed often to actually compute determinants. A particularly interesting consequence is to note what it means for existence of inverses. Take A and B to be inverses, that is $AB = I$. Then

$$\det(A) \det(B) = \det(AB) = \det(I) = 1.$$

Neither $\det(A)$ nor $\det(B)$ can be zero. This fact is an extremely useful property of the determinant, and one which is used often in this book:

Theorem 3.5.3

[thm:detInv] An $n \times n$ matrix A is invertible if and only if $\det(A) \neq 0$.

In fact, $\det(A^{-1}) \det(A) = 1$ says that

$$\det(A^{-1}) = \frac{1}{\det(A)}.$$

So we know what the determinant of A^{-1} is without computing A^{-1} .

Let us return to the formula for the inverse of a 2×2 matrix:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Notice the determinant of the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ in the denominator of the fraction. The formula only works if the determinant is nonzero, otherwise we are dividing by zero.

A common notation for the determinant is a pair of vertical lines:

$$\left| \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right| = \det \left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \right).$$

Personally, I find this notation confusing as vertical lines usually mean a positive quantity, while determinants can be negative. Also think about how to write the absolute value of a determinant. This notation is not used in this book.

With this discussion of determinants complete, we can now state a major theorem from linear algebra that will help us here and when we get back to solving differential equations using this linear algebra. In a full course on linear algebra, this theorem would be covered in full detail, including all of the proofs. For this introduction, we give some idea as to why everything is true here, but not all of the details.

Note: This is an example of an *equivalence* theorem, which is fairly common in mathematics. It means that if any one of the statements are true, then we know that all of the others are true as well. It means it's harder to prove, but once we have such a theorem, it is very powerful in how we can use it going forward.

Theorem 3.5.4

Let A be an $n \times n$ matrix. The following are equivalent:

- (a) A is invertible.
- (b) $\det(A) \neq 0$.
- (c) There is a unique solution to $A\vec{x} = \vec{b}$ for every vector \vec{b} .
- (d) The only solution to $A\vec{x} = \vec{0}$ is $\vec{x} = \vec{0}$.
- (e) The reduced row echelon form of A is I_n , the identity matrix.
- (f) The rank of A is n .
- (g) The rows of A are linearly independent.
- (h) The columns of A are linearly independent.

Proof. Why is all of this true? For (a) and (b), we have Theorem ?? to say that they are equivalent. For (c), if A is invertible, then the unique solution to $A\vec{x} = \vec{b}$ is $\vec{x} = A^{-1}\vec{b}$. If we take $\vec{b} = \vec{0}$ here, we get (d), that the solution is $\vec{x} = A^{-1}\vec{0} = \vec{0}$. This means that reducing the system of equations $A\vec{x} = 0$ gives $x_1 = 0, x_2 = 0, \dots, x_n = 0$, which means the reduced row echelon form of A is just the identity matrix, which is (e). This has n pivot rows, so that the rank of A is n . Finally, this means that the dimension of the column space and row space is both n , and since there are n of these vectors, it means they are all linearly independent. \square

This is a massive theorem that forms most of the backbone of linear algebra. We will only be using a few parts of it later, but since we have seen all of the components, it is nice to see them all put together into one complete statement.

3.5.1 Exercises

Exercise 3.5.2: Compute the determinant of the following matrices:

a) $[3]$	b) $\begin{bmatrix} 1 & 3 \\ 2 & 1 \end{bmatrix}$	c) $\begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix}$	d) $\begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}$
e) $\begin{bmatrix} 2 & 1 & 0 \\ -2 & 7 & -3 \\ 0 & 2 & 0 \end{bmatrix}$	f) $\begin{bmatrix} 2 & 1 & 3 \\ 8 & 6 & 3 \\ 7 & 9 & 7 \end{bmatrix}$	g) $\begin{bmatrix} 0 & 2 & 5 & 7 \\ 0 & 0 & 2 & -3 \\ 3 & 4 & 5 & 7 \\ 0 & 0 & 2 & 4 \end{bmatrix}$	h) $\begin{bmatrix} 0 & 1 & 2 & 0 \\ 1 & 1 & -1 & 2 \\ 1 & 1 & 2 & 1 \\ 2 & -1 & -2 & 3 \end{bmatrix}$

Exercise 3.5.3:* Compute the determinant of the following matrices:

$$\begin{array}{llll}
 a) [-2] & b) \begin{bmatrix} 2 & -2 \\ 1 & 3 \end{bmatrix} & c) \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} & d) \begin{bmatrix} 2 & 9 & -11 \\ 0 & -1 & 5 \\ 0 & 0 & 3 \end{bmatrix} \\
 e) \begin{bmatrix} 2 & 1 & 0 \\ -2 & 7 & 3 \\ 1 & 1 & 0 \end{bmatrix} & f) \begin{bmatrix} 5 & 1 & 3 \\ 4 & 1 & 1 \\ 4 & 5 & 1 \end{bmatrix} & g) \begin{bmatrix} 3 & 2 & 5 & 7 \\ 0 & 0 & 2 & 0 \\ 0 & 4 & 5 & 0 \\ 2 & 1 & 2 & 4 \end{bmatrix} & h) \begin{bmatrix} 0 & 2 & 1 & 0 \\ 1 & 2 & -3 & 4 \\ 5 & 6 & -7 & 8 \\ 1 & 2 & 3 & -2 \end{bmatrix}
 \end{array}$$

Exercise 3.5.4: For which x are the following matrices singular (not invertible).

$$\begin{array}{llll}
 a) \begin{bmatrix} 2 & 3 \\ 2 & x \end{bmatrix} & b) \begin{bmatrix} 2 & x \\ 1 & 2 \end{bmatrix} & c) \begin{bmatrix} x & 1 \\ 4 & x \end{bmatrix} & d) \begin{bmatrix} x & 0 & 1 \\ 1 & 4 & 2 \\ 1 & 6 & 2 \end{bmatrix}
 \end{array}$$

Exercise 3.5.5:* For which x are the following matrices singular (not invertible).

$$\begin{array}{llll}
 a) \begin{bmatrix} 1 & 3 \\ 1 & x \end{bmatrix} & b) \begin{bmatrix} 3 & x \\ 1 & 3 \end{bmatrix} & c) \begin{bmatrix} x & 3 \\ 3 & x \end{bmatrix} & d) \begin{bmatrix} x & 1 & 0 \\ 1 & 4 & 0 \\ 1 & 6 & 2 \end{bmatrix}
 \end{array}$$

Exercise 3.5.6: Is the matrix A below invertible? How do you know?

$$A = \begin{bmatrix} 4 & 0 & 3 & 1 \\ 2 & 1 & -2 & 0 \\ 0 & 0 & 1 & -3 \\ 3 & 2 & 1 & -5 \end{bmatrix}$$

Exercise 3.5.7:* Compute the rank of the matrix A below.

$$A = \begin{bmatrix} 0 & -3 & 2 & 4 \\ -5 & -4 & -5 & -1 \\ 1 & 4 & -3 & -5 \\ -2 & -3 & -2 & 1 \end{bmatrix}$$

What does this tell you about the invertibility of A ? How about the solutions to $A\vec{x} = \vec{0}$?

Exercise 3.5.8:* Compute the rank of the matrix A below.

$$A = \begin{bmatrix} 3 & -5 & 5 \\ 2 & -3 & 3 \\ 4 & 0 & -1 \end{bmatrix}$$

What does this tell you about the invertibility of A ? How about the solutions to $A\vec{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$?

Exercise 3.5.9:* Compute the determinant of the matrix

$$\begin{bmatrix} 5 & 4 & 3 \\ -4 & -3 & -4 \\ -5 & -5 & 4 \end{bmatrix}$$

using row reduction.

Exercise 3.5.10:* Compute the determinant of the matrix

$$\begin{bmatrix} -5 & -3 & -5 & -1 \\ 4 & 0 & -5 & 4 \\ 0 & -2 & -1 & -2 \\ -1 & -5 & -4 & -4 \end{bmatrix}$$

using row reduction.

Exercise 3.5.11:* Compute the determinant of the matrix

$$\begin{bmatrix} 4 & 1 & -3 & 0 \\ -1 & 4 & 2 & -2 \\ -1 & -3 & 3 & 2 \\ -5 & -4 & 1 & 1 \end{bmatrix}$$

using row reduction.

Exercise 3.5.12: Compute

$$\det \left(\begin{bmatrix} 2 & 1 & 2 & 3 \\ 0 & 8 & 6 & 5 \\ 0 & 0 & 3 & 9 \\ 0 & 0 & 0 & 1 \end{bmatrix}^{-1} \right)$$

without computing the inverse.

Exercise 3.5.13:* Compute

$$\det \left(\begin{bmatrix} 3 & 4 & 7 & 12 \\ 0 & -1 & 9 & -8 \\ 0 & 0 & -2 & 4 \\ 0 & 0 & 0 & 2 \end{bmatrix}^{-1} \right)$$

without computing the inverse.

Exercise 3.5.14: Suppose

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 7 & \pi & 1 & 0 \\ 2^8 & 5 & -99 & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 5 & 9 & 1 & -\sin(1) \\ 0 & 1 & 88 & -1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Let $A = LU$. Compute $\det(A)$ in a simple way, without computing what is A . Hint: First read off $\det(L)$ and $\det(U)$.

Exercise 3.5.15: Consider the linear mapping from \mathbb{R}^2 to \mathbb{R}^2 given by the matrix $A = \begin{bmatrix} 1 & x \\ 2 & 1 \end{bmatrix}$ for some number x . You wish to make A such that it doubles the area of every geometric figure. What are the possibilities for x (there are two answers).

Exercise 3.5.16 (challenging):* Find all the x that make the matrix inverse

$$\begin{bmatrix} 1 & 2 \\ 1 & x \end{bmatrix}^{-1}$$

have only integer entries (no fractions). Note that there are two answers.

Exercise 3.5.17: Suppose A and S are $n \times n$ matrices, and S is invertible. Suppose that $\det(A) = 3$. Compute $\det(S^{-1}AS)$ and $\det(SAS^{-1})$. Justify your answer using the theorems in this section.

Exercise 3.5.18: Let A be an $n \times n$ matrix such that $\det(A) = 1$. Compute $\det(xA)$ given a number x . Hint: First try computing $\det(xI)$, then note that $xA = (xI)A$.

3.6 Eigenvalues and Eigenvectors

Learning Objectives

After this section, you will be able to:

- Find the eigenvalues and eigenvectors of a matrix,
- Use complex numbers to find eigenvalues and eigenvectors if necessary, and
- Identify the algebraic and geometric multiplicity of an eigenvalue to determine if it is defective.

Consider the matrix

$$A = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix}.$$

We can compute a few operations with this matrix. For instance

$$A \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

and

$$A \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \end{bmatrix}.$$

This last computation is fairly interesting, because the result we get is the same as 3 times the original vector. However, the matrix A does not multiply every vector by 3, as seen in the first example and the fact that

$$A \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

so A actually preserves this vector, multiplying it by 1. So, these vectors, $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$, and numbers, 3 and 1, are somehow special for this matrix A . With this information, we want to define these vectors as *eigenvectors* and numbers as *eigenvalues* of the matrix A .

Definition 3.6.1

For a square matrix A , we say that non-zero vector \vec{v} is an *eigenvector* of the matrix A if there exists a number λ so that

$$A\vec{v} = \lambda\vec{v}.$$

In this case, we say that λ is an *eigenvalue* of A and it is the *corresponding eigenvalue* for the eigenvector \vec{v} .

Thus, we can say that, for the matrix

$$A = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix},$$

we see that $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$ is an eigenvector with corresponding eigenvalue 3, and that $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$ is an eigenvector with corresponding eigenvalue 1.

Why are these important? It turns out that these eigenvalues and eigenvectors characterize the behavior of the matrix A . For example, if we wanted to figure out what happens when A is applied to the vector $\begin{bmatrix} 6 \\ 4 \end{bmatrix}$, we can figure this out as

$$\begin{aligned} A \begin{bmatrix} 6 \\ 4 \end{bmatrix} &= A \left(\begin{bmatrix} 4 \\ 3 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix} \right) \\ &= A \begin{bmatrix} 4 \\ 3 \end{bmatrix} + A \begin{bmatrix} 2 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 4 \\ 3 \end{bmatrix} + 3 \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 10 \\ 6 \end{bmatrix} \end{aligned}$$

In addition, eigenvectors determine directions in which multiplying by the matrix A behaves just like scalar multiplication. This idea will be very important for our understanding of systems of differential equations, because we have already seen how to solve a scalar first order equation way back in § 0.1 and § 1.3.

3.6.1 Finding Eigenvalues and Eigenvectors

Since eigenvalues and eigenvectors are so important, we want to know how to find them. To do this, we are looking for a number λ and a non-zero vector \vec{v} so that

$$A\vec{v} = \lambda\vec{v}.$$

We can rewrite this as

$$A\vec{v} - \lambda\vec{v} = 0$$

or, using the identity matrix,

$$(A - \lambda I)\vec{v} = 0.$$

This means that we are looking for a non-zero solution to a homogeneous vector equation of the form $B\vec{v} = 0$. This is where all of our linear algebra theory comes into play.

Theorem 3.5.4 tells us that, combining parts (b) and (d), that there is a non-zero solution to $(A - \lambda I)\vec{v} = 0$ if and only if the determinant of the matrix $A - \lambda I$ is zero. Therefore, we can compute this determinant, find the values of λ so that $\det(A - \lambda I) = 0$, and these will give us our eigenvalues. Let's see an example of what this looks like.

Example 3.6.1: Compute $\det(A - \lambda I)$ for the matrix

$$A = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix}.$$

Solution: For this matrix, we have that

$$A - \lambda I = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 7 - \lambda & -8 \\ 3 & -3 - \lambda \end{bmatrix}.$$

Thus

$$\begin{aligned}\det(A - \lambda I) &= \det \left(\begin{bmatrix} 7 - \lambda & -8 \\ 3 & -3 - \lambda \end{bmatrix} \right) \\ &= (7 - \lambda)(-3 - \lambda) - (-8)(3) = \lambda^2 + 3\lambda - 7\lambda - 21 + 24 \\ &= \lambda^2 - 4\lambda + 3\end{aligned}$$

If we were looking for eigenvalues here, we could then set this equal to zero, getting that

$$0 = \lambda^2 - 4\lambda + 3 = (\lambda - 1)(\lambda - 3)$$

so that the eigenvalues are 1 and 3. □

In this case, we saw that computing $\det(A - \lambda I)$ for this case, we ended up with a quadratic polynomial, so it was easy to find the eigenvalues. Thankfully, no matter the size of the matrix, we will always get a polynomial here. For a matrix A , the expression $\det(A - \lambda I)$ is called the *characteristic polynomial* of the matrix. It will always be a polynomial, and for A an $n \times n$ matrix, it will be a degree n polynomial. This explains why we got a quadratic polynomial for the 2×2 matrix A . Therefore, for a matrix A , the roots of the characteristic polynomial are the eigenvalues of A .

Once we have the eigenvalues, we can use them to find the eigenvectors. As with how we started this discussion, we are looking for a non-zero vector \vec{v} so that

$$(A - \lambda I)\vec{v} = 0,$$

and we know the value of λ . Therefore, we can set up a system of equations that corresponds to

$$(A - \lambda I)\vec{v} = 0$$

and solve it for the components of the eigenvector.

Example 3.6.2: Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix}.$$

Solution: The previous example shows that the eigenvalues for this matrix are 1 and 3. For the eigenvalue 1, we want to find a non-zero solution to $(A - I)\vec{v} = 0$, which means we want to solve for

$$(A - I)\vec{v} = \begin{bmatrix} 7 - 1 & -8 \\ 3 & -3 - 1 \end{bmatrix} \vec{v} = \begin{bmatrix} 6 & -8 \\ 3 & -4 \end{bmatrix} \vec{v} = 0.$$

Writing the vector \vec{v} as $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, this system of equations becomes

$$\begin{aligned}6v_1 - 8v_2 &= 0 \\ 3v_1 - 4v_2 &= 0\end{aligned}$$

Since the second equation is two times the first one, these equations are redundant, so we only need to satisfy $3v_1 - 4v_2 = 0$. We can do this by choosing $v_1 = 4$ and $v_2 = 3$, which gives that for $\lambda = 1$, a corresponding eigenvector is $\begin{bmatrix} 4 \\ 3 \end{bmatrix}$.

We can follow the same process for the eigenvalue 3. For this, we want to find a non-zero solution to $(A - 3I)\vec{v} = 0$, which means that we want to solve

$$(A - 3I)\vec{v} = \begin{bmatrix} 7 - 3 & -8 \\ 3 & -3 - 3 \end{bmatrix} \vec{v} = \begin{bmatrix} 4 & -8 \\ 3 & -6 \end{bmatrix} \vec{v} = 0.$$

Writing the vector \vec{v} as $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$, we get the two equations

$$\begin{aligned} 4v_1 - 8v_2 &= 0 \\ 3v_1 - 6v_2 &= 0 \end{aligned}$$

As before, these two equations are the same, since they are both a multiple of $v_1 - 2v_2 = 0$. Therefore, we just need to find a solution to that previous equation, which can be done with $v_1 = 2$ and $v_2 = 1$. Therefore, an eigenvector for eigenvalue 3 is $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$. □

This example illustrates the standard process that is always used to find eigenvalues and eigenvectors of matrices: find the characteristic polynomial, get the roots of this polynomial, and use each of these eigenvalues to set up a system of equations for the components of each eigenvector. In addition, the equations that we get from this system will always be redundant if we have found the eigenvalue correctly. Since $\det(A - \lambda I) = 0$, we know that the rows of the matrix $A - \lambda I$ are not linearly independent, and so the row-echelon form of $A - \lambda I$ must have a zero row in it. This process works for any size matrix, but it becomes harder to find the roots of this polynomial when it is higher degree.

Example 3.6.3: Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & 6 & 0 \\ 9 & -4 & 10 \\ 2 & -6 & 3 \end{bmatrix}.$$

Solution: We start by hunting for eigenvalues by taking the determinant of $A - \lambda I$, which will require the cofactor expansion in order to solve.

$$\begin{aligned}
\det(A - \lambda I) &= \det \left(\begin{bmatrix} 1 - \lambda & 6 & 0 \\ 9 & -4 - \lambda & 10 \\ 2 & -6 & 3 - \lambda \end{bmatrix} \right) \\
&= (1 - \lambda) \det \left(\begin{bmatrix} -4 - \lambda & 10 \\ -6 & 3 - \lambda \end{bmatrix} \right) - 6 \det \left(\begin{bmatrix} 9 & 10 \\ 2 & 3 - \lambda \end{bmatrix} \right) \\
&= (1 - \lambda)((-4 - \lambda)(3 - \lambda) + 60) - 6(9(3 - \lambda) - 20) \\
&= (1 - \lambda)(\lambda^2 + 4\lambda - 3\lambda - 12 + 60) - 6(27 - 9\lambda - 20) \\
&= (1 - \lambda)(\lambda^2 + \lambda + 48) - 42 + 54\lambda \\
&= \lambda^2 + \lambda + 48 - \lambda^3 - \lambda^2 - 48\lambda - 42 + 54\lambda \\
&= -\lambda^3 + 7\lambda + 6
\end{aligned}$$

We need to look for the roots of this polynomial. There's no nice way to factor this right away, so we need to start guessing roots. We know that the root must be a factor of 6. If we try $\lambda = 1$, we get

$$-1 + 7 + 6 = 12 \neq 0$$

so that one doesn't work. Plugging in $\lambda = -1$, we get

$$-(-1)^3 - 7 + 6 = 1 - 7 + 6 = 0$$

so this is a root, meaning that $\lambda + 1$ is a factor of the characteristic polynomial. We can then use polynomial long division to get that

$$-\lambda^3 + 7\lambda + 6 = (\lambda + 1)(-\lambda^2 + \lambda + 6) = -(\lambda + 1)(\lambda^2 - \lambda - 6)$$

and the quadratic term here factors as $(\lambda - 3)(\lambda + 2)$. Thus, the characteristic polynomial of this matrix is

$$(\lambda + 1)(\lambda - 3)(\lambda + 2)$$

so the eigenvalues are -1 , 3 , and -2 .

For the eigenvalue -1 , the eigenvector must satisfy

$$(A + I)\vec{v} = \vec{0}$$

which we can write as

$$\begin{bmatrix} 2 & 6 & 0 \\ 9 & -3 & 10 \\ 2 & -6 & 4 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \vec{0}.$$

To solve this, we row-reduce the coefficient matrix.

$$\begin{aligned} \left[\begin{array}{ccc} 2 & 6 & 0 \\ 9 & -3 & 10 \\ 2 & -6 & 4 \end{array} \right] &\rightarrow \left[\begin{array}{ccc} 1 & 3 & 0 \\ 9 & -3 & 10 \\ 2 & -6 & 4 \end{array} \right] \\ &\rightarrow \left[\begin{array}{ccc} 1 & 3 & 0 \\ 0 & -30 & 10 \\ 0 & -12 & 4 \end{array} \right] \\ &\rightarrow \left[\begin{array}{ccc} 1 & 3 & 0 \\ 0 & -3 & 1 \\ 0 & -12 & 4 \end{array} \right] \\ &\rightarrow \left[\begin{array}{ccc} 1 & 3 & 0 \\ 0 & -3 & 1 \\ 0 & 0 & 0 \end{array} \right] \end{aligned}$$

Therefore, the eigenvector must satisfy $v_1 + 3v_2 = 0$ and $-3v_2 + v_3 = 0$. We need to pick any non-zero set of numbers that solves these equations. For example, we could pick $v_2 = 1$ to get that we need $v_1 = -3$ and $v_3 = 3$. This gives an eigenvector of

$$\begin{bmatrix} -3 \\ 1 \\ 3 \end{bmatrix}.$$

For the eigenvalue 3, the eigenvector must satisfy

$$\left[\begin{array}{ccc} -2 & 6 & 0 \\ 9 & -7 & 10 \\ 2 & -6 & 0 \end{array} \right] \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \vec{0}.$$

Row reduction gives

$$\begin{aligned} \left[\begin{array}{ccc} -2 & 6 & 0 \\ 9 & -7 & 10 \\ 2 & -6 & 0 \end{array} \right] &\rightarrow \left[\begin{array}{ccc} 1 & -3 & 0 \\ 9 & -7 & 10 \\ 2 & -6 & 0 \end{array} \right] \\ &\rightarrow \left[\begin{array}{ccc} 1 & -3 & 0 \\ 0 & 20 & 10 \\ 0 & 0 & 0 \end{array} \right] \\ &\rightarrow \left[\begin{array}{ccc} 1 & -3 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 0 \end{array} \right] \end{aligned}$$

which means that the eigenvector must satisfy $v_1 - 3v_2 = 0$ and $2v_2 + v_3 = 0$. Again, choosing $v_2 = 1$ gives that we want $v_1 = 3$ and $v_3 = -2$. Therefore, a corresponding eigenvector here is

$$\begin{bmatrix} 3 \\ 1 \\ -2 \end{bmatrix}.$$

For the eigenvalue -2 , the eigenvector must satisfy

$$\begin{bmatrix} 3 & 6 & 0 \\ 9 & -2 & 10 \\ 2 & -6 & 5 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \vec{0}$$

where we can row reduce the coefficient matrix.

$$\begin{aligned} \begin{bmatrix} 3 & 6 & 0 \\ 9 & -2 & 10 \\ 2 & -6 & 5 \end{bmatrix} &\rightarrow \begin{bmatrix} 1 & 2 & 0 \\ 9 & -2 & 10 \\ 2 & -6 & 5 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & 2 & 0 \\ 0 & -20 & 10 \\ 0 & -10 & 5 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & 2 & 0 \\ 0 & -2 & 1 \\ 0 & -10 & 5 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & 2 & 0 \\ 0 & -2 & 1 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Therefore, the eigenvector must satisfy $v_1 + 2v_2 = 0$ and $-2v_2 + v_3 = 0$. Picking $v_2 = 1$ again gives that we want $v_1 = -2$ and $v_3 = 2$. Therefore, an eigenvector with eigenvalue -2 is

$$\begin{bmatrix} -2 \\ 1 \\ 2 \end{bmatrix}.$$

□

3.6.2 Real Eigenvalues

Since eigenvalues come from finding the roots of a polynomial, there are a few different situations that can arise in terms of these eigenvalues. If we take a quadratic polynomial, there are three options for the two roots.

- Two real and different roots,
- Two complex roots in a conjugate pair, or
- One double (repeated) root.

The same is true for eigenvalues, they are either all real and distinct, there are some that appear in complex conjugate pairs, or there are some repeated eigenvalues. The easiest of these cases is when the characteristic polynomial has all real and distinct eigenvalues.

In this case, we get a very nice result. We know that for each eigenvalue, there will always be at least one eigenvector, otherwise it wouldn't be an eigenvalue. If the matrix A is an

$n \times n$ matrix, then the characteristic polynomial is a degree n polynomial, which will have n distinct roots by our assumption. Each of these will have a corresponding eigenvector, giving us n eigenvectors as well. A more involved result tells us that eigenvectors for different eigenvalues are always linearly independent. Therefore, we get n vectors in \mathbb{R}^n , that are linearly independent, and so they are a basis. This gives the following result.

Theorem 3.6.1

Let A be an $n \times n$ matrix. Assume that the characteristic polynomial of A has all real and distinct roots, namely that

$$\det(A - \lambda I) = (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n)$$

for $\lambda_1, \dots, \lambda_n$ the distinct real eigenvalues. Then there exist vectors $\vec{v}_1, \dots, \vec{v}_n$ such that \vec{v}_i is an eigenvector for eigenvalue λ_i and $\{\vec{v}_1, \dots, \vec{v}_n\}$ form a basis of \mathbb{R}^n .

This is useful to know for now, but will be critical when we want to use this information to solve systems of differential equations later.

3.6.3 Complex Eigenvalues

When the matrix has complex eigenvalues, the process is very similar to before. However, the eigenvector will necessarily also be complex, that is, some of the components of this vector will be complex numbers. Let's illustrate this with an example.

Example 3.6.4: Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 3 & -8 \\ 5 & -9 \end{bmatrix}.$$

Solution: We first look for the eigenvalues using the characteristic polynomial of A .

$$\begin{aligned} \det(A - \lambda I) &= \det \left(\begin{bmatrix} 3 - \lambda & -8 \\ 5 & -9 - \lambda \end{bmatrix} \right) \\ &= (3 - \lambda)(-9 - \lambda) + 40 \\ &= \lambda^2 + 9\lambda - 3\lambda - 27 + 40 \\ &= \lambda^2 + 6\lambda + 13 \end{aligned}$$

This quadratic does not factor, so we use the quadratic formula to find that

$$\lambda = \frac{-6 \pm \sqrt{6^2 - 4 \cdot 13}}{2} = \frac{-6 \pm \sqrt{-16}}{2} = -3 \pm 2i$$

so that we have complex eigenvalues.

We now look for the eigenvectors in the same way as in the real case. If we take the eigenvalue $-3 + 2i$, then such an eigenvector must satisfy

$$(A - (-3 + 2i)I)\vec{v} = \vec{0}.$$

This means that

$$\begin{bmatrix} 3 - (-3 + 2i) & -8 \\ 5 & -9 - (-3 + 2i) \end{bmatrix} \vec{v} = \begin{bmatrix} 6 - 2i & -8 \\ 5 & -6 - 2i \end{bmatrix} \vec{v} = \vec{0}.$$

These two equations should be redundant, and to verify that, we will multiply the top row by $6 + 2i$ in row reduction to get

$$\begin{aligned} \begin{bmatrix} 6 - 2i & -8 \\ 5 & -6 - 2i \end{bmatrix} &\rightarrow \begin{bmatrix} (6 - 2i)(6 + 2i) & -8(6 + 2i) \\ 5 & -6 - 2i \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 40 & -48 - 16i \\ 5 & -6 - 2i \end{bmatrix} \end{aligned}$$

and from this, we can see that the top row is 8 times the bottom one, so they are redundant. Thus, an eigenvector must satisfy

$$5v_1 - (6 + 2i)v_2 = 0$$

and we can pick any non-zero numbers that satisfy this. One simple way to do this is by switching the coefficients, so that $v_1 = 6 + 2i$ and $v_2 = 5$. Therefore, an eigenvector that we get is

$$\begin{bmatrix} 6 + 2i \\ 5 \end{bmatrix}.$$

Now, we can take the other eigenvalue, $-3 - 2i$. The process is the same, so that the vector must satisfy

$$\begin{bmatrix} 3 - (-3 - 2i) & -8 \\ 5 & -9 - (-3 - 2i) \end{bmatrix} \vec{v} = \begin{bmatrix} 6 + 2i & -8 \\ 5 & -6 + 2i \end{bmatrix} \vec{v} = \vec{0}.$$

To check redundancy again, we multiply the top row by $6 - 2i$ to get

$$\begin{aligned} \begin{bmatrix} 6 + 2i & -8 \\ 5 & -6 + 2i \end{bmatrix} &\rightarrow \begin{bmatrix} (6 + 2i)(6 - 2i) & -8(6 - 2i) \\ 5 & -6 + 2i \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 40 & -48 + 16i \\ 5 & -6 + 2i \end{bmatrix} \end{aligned}$$

and again, the first equation is 8 times the second one. Thus, the eigenvector will need to satisfy

$$5v_1 - (6 - 2i)v_2 = 0$$

which can be done by picking $v_1 = 6 - 2i$ and $v_2 = 5$, giving an eigenvector of

$$\begin{bmatrix} 6 - 2i \\ 5 \end{bmatrix}.$$

\(\square\)

The process here is the same as it was in the real case, except that now all of the equations are complex equations. In particular, the “redundancy” that we expect to see between the

equations will likely be via a complex multiple. The easiest way to verify that these equations are redundant is by multiplying the first entry in each row by its complex conjugate. This is because, if we have the complex number $a + bi$, multiplying this by $a - bi$ gives

$$(a + bi)(a - bi) = a^2 + abi - abi - b^2i^2 = a^2 + b^2$$

which is now a real number. This will make it easier to compare the two equations to make sure that they are redundant, and that the eigenvalue was found correctly.

Example 3.6.5: Find the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 1 & 9 & 6 \\ 0 & 1 & 6 \\ 0 & -3 & -5 \end{bmatrix}.$$

Solution: We first look for eigenvalues, like always. We get these by computing

$$\det(A - \lambda I) = \det \left(\begin{bmatrix} 1 - \lambda & 9 & 6 \\ 0 & 1 - \lambda & 6 \\ 0 & -3 & -5 - \lambda \end{bmatrix} \right).$$

We will compute this by cofactor expansion along the second row.

$$\begin{aligned} \det \left(\begin{bmatrix} 1 - \lambda & 9 & 6 \\ 0 & 1 - \lambda & 6 \\ 0 & -3 & -5 - \lambda \end{bmatrix} \right) &= (-1)^{2+2}(1 - \lambda) \det \left(\begin{bmatrix} 1 - \lambda & 6 \\ 0 & -5 - \lambda \end{bmatrix} \right) \\ &\quad + (-1)^{2+3}6 \det \left(\begin{bmatrix} 1 - \lambda & 9 \\ 0 & -3 \end{bmatrix} \right) \\ &= (1 - \lambda)(1 - \lambda)(-5 - \lambda) - 6(1 - \lambda)(-3) \\ &= (1 - \lambda)((1 - \lambda)(-5 - \lambda) + 18) \\ &= (1 - \lambda)(\lambda^2 + 4\lambda + 13) \end{aligned}$$

so that one eigenvalue is at $\lambda = 1$. For the other two, we use the quadratic formula to obtain

$$\lambda = \frac{-4 \pm \sqrt{16 - 4 \cdot 13}}{2} = \frac{-4 \pm \sqrt{-36}}{2} = -2 \pm 3i.$$

Thus, we have one real eigenvalue and two complex eigenvalues.

For $\lambda = 1$, we know that the eigenvector must satisfy

$$\begin{bmatrix} 0 & 9 & 6 \\ 0 & 0 & 6 \\ 0 & -3 & -6 \end{bmatrix} \vec{v} = \vec{0}.$$

Row reduction will reduce this matrix to

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

(Check this!) so that the eigenvector in this case is

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

For the eigenvalue $-2 + 3i$, we get that the eigenvector must satisfy

$$\begin{bmatrix} 3 - 3i & 9 & 6 \\ 0 & 3 - 3i & 6 \\ 0 & -3 & -3 - 3i \end{bmatrix} \vec{v} = \vec{0}.$$

We now want to row reduce the coefficient matrix. To do so, we start by dividing the first row by 3 then multiplying by $1 + i$.

$$\begin{aligned} \begin{bmatrix} 3 - 3i & 9 & 6 \\ 0 & 3 - 3i & 6 \\ 0 & -3 & -3 - 3i \end{bmatrix} &\rightarrow \begin{bmatrix} 1 - i & 3 & 2 \\ 0 & 3 - 3i & 6 \\ 0 & -3 & -3 - 3i \end{bmatrix} \\ &\rightarrow \begin{bmatrix} (1 - i)(1 + i) & 3(1 + i) & 2(1 + i) \\ 0 & 3 - 3i & 6 \\ 0 & -3 & -3 - 3i \end{bmatrix}. \\ &\rightarrow \begin{bmatrix} 2 & 3 + 3i & 2 + 2i \\ 0 & 3 - 3i & 6 \\ 0 & -3 & -3 - 3i \end{bmatrix} \end{aligned}$$

We could divide the first row by 2 to get to a 1 in the top-right entry, but we'll wait on that in order to avoid fractions. To row reduce the rest of the matrix, we will divide each of the remaining two rows by 3, and then multiply the second by $1 + i$, just like we did to the first row.

$$\begin{aligned} \begin{bmatrix} 2 & 3 + 3i & 2 + 2i \\ 0 & 3 - 3i & 6 \\ 0 & -3 & -3 - 3i \end{bmatrix} &\rightarrow \begin{bmatrix} 2 & 3 + 3i & 2 + 2i \\ 0 & 1 - i & 2 \\ 0 & -1 & -1 - i \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 2 & 3 + 3i & 2 + 2i \\ 0 & 2 & 2 + 2i \\ 0 & -1 & -1 - i \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 2 & 3 + 3i & 2 + 2i \\ 0 & 1 & 1 + i \\ 0 & -1 & -1 - i \end{bmatrix} \end{aligned}$$

which illustrates that the last two rows are redundant. Thus, the reduced form of the matrix that we have (which is not quite a row echelon form, but it is enough to back-solve for the eigenvector) is

$$\begin{bmatrix} 2 & 3 + 3i & 2 + 2i \\ 0 & 1 & 1 + i \\ 0 & 0 & 0 \end{bmatrix}.$$

This means that the eigenvector \vec{v} must satisfy

$$2v_1 + (3 + 3i)v_2 + (2 + 2i)v_3 = 0 \quad v_2 + (1 + i)v_3 = 0.$$

We can satisfy the second of these equations by choosing $v_2 = 1 + i$ and $v_3 = -1$. Plugging these values into the first equation gives that

$$\begin{aligned} 0 &= 2v_1 + (3 + 3i)v_2 + (2 + 2i)v_3 \\ &= 2v_1 + (3 + 3i)(1 + i) + (2 + 2i)(-1) \\ &= 2v_1 + 3 + 3i + 3i - 3 - 2 - 2i \\ &= 2v_1 - 2 + 4i \end{aligned}$$

Therefore, we need to take $v_1 = 1 - 2i$, giving that the eigenvector is

$$\begin{bmatrix} 1 - 2i \\ 1 + i \\ -1 \end{bmatrix}.$$

A very similar computation following the same set of steps (or just using the remark below) for the eigenvalue $-2 - 3i$ gives that this corresponding eigenvector is

$$\begin{bmatrix} 1 + 2i \\ 1 - i \\ -1 \end{bmatrix}.$$

\(\square\)

One fact that comes out of those examples is that the eigenvectors for conjugate eigenvalues are also complex conjugates. This comes from the fact that A is a real matrix, which means that if

$$A\vec{v} = \lambda\vec{v}$$

and we take the complex conjugate of both sides, we get that

$$A\bar{\vec{v}} = \bar{A}\vec{v} = \bar{\lambda}\vec{v} = \bar{\lambda}\bar{\vec{v}}$$

so that $\bar{\vec{v}}$ is an eigenvector for $\bar{\lambda}$. This means that, when solving these types of problems, we only need to find one of the complex eigenvectors and can get the other by taking the complex conjugate.

3.6.4 Repeated Eigenvalues

Distinct and complex eigenvalues all work out nicely and in pretty much the same manner. For repeated eigenvalues, the issues get more significant.

Example 3.6.6: Find the eigenvalues and eigenvectors of the matrices

$$A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \quad B = \begin{bmatrix} 4 & -1 \\ 1 & 2 \end{bmatrix}.$$

Solution: For the matrix A , we can compute the characteristic polynomial

$$\det(A - \lambda I) = \det \begin{pmatrix} 3 - \lambda & 0 \\ 0 & 3 - \lambda \end{pmatrix} = (3 - \lambda)(3 - \lambda)$$

Therefore, we have a double root at 3 for this matrix. Therefore, the only eigenvalue we get is 3. When we look to find the eigenvectors, we get

$$A - 3I = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

so that this matrix multiplied by *any* vector is zero. Therefore, we can use both $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ as eigenvectors.

On the other hand, the matrix B has a characteristic polynomial

$$\begin{aligned} \det(B - \lambda I) &= \det \begin{pmatrix} 4 - \lambda & -1 \\ 1 & 2 - \lambda \end{pmatrix} \\ &= (4 - \lambda)(2 - \lambda) - (-1)(1) = \lambda^2 - 6\lambda + 8 + 1 \\ &= \lambda^2 - 6\lambda + 9 = (\lambda - 3)^2 \end{aligned}$$

so again, we have a double root at 3. However, when we go to find the eigenvectors, we get that

$$B - 3I = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$$

which gives that an eigenvector must satisfy $v_1 - v_2 = 0$ so $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ works. □

There is a big difference between these two examples. Both had the same characteristic polynomial of $(\lambda - 3)^2$, but for A , we could find two linearly independent eigenvectors, but for B , we could only find 1. This seems like it might be a problem, since we would like to get to two eigenvectors like we did for both of the previous two cases. This leads us to define the following for A and $n \times n$ matrix and r an eigenvalue of A .

Definition 3.6.2

- The *algebraic multiplicity* of r is the power of $(\lambda - r)$ in the characteristic polynomial if A .
- The *geometric multiplicity* of r is the number of linearly independent eigenvectors of A with eigenvalue r .
- The *defect* of r is the difference between the algebraic multiplicity and the geometric multiplicity of r .
- We say that an eigenvalue is *defective* if the defect is at least 1.

For the previous example, the algebraic multiplicity of 3 for both A and B was 2, but the geometric multiplicity of 3 for A is 2, and for B is it only 1. Therefore A has a defect of 0 and B has a defect of 1, so 3 is a defective eigenvalue for matrix B .

In terms of these multiplicities, there are two facts that are known to be true.

1. If r is an eigenvalue, then both the algebraic and geometric multiplicity are at least 1.
2. The algebraic multiplicity of any eigenvalue is always greater than or equal to the geometric multiplicity.

This tells us that in the case of real and distinct eigenvalues, every eigenvalue has multiplicity 1. Since the geometric multiplicity is also 1, this means that none of these eigenvalues are defective. This was great, because it let us get to n eigenvectors for an $n \times n$ matrix, and these generated a basis of \mathbb{R}^n .

Why is a defective eigenvalue a problem? When we go solve differential equations using the method in Chapter 4, having a ‘full set’ of eigenvectors, or n eigenvectors for an $n \times n$ matrix, will be very important. When we have a defective eigenvalue, we can’t get there. Since the degree of the characteristic polynomial is n , the only way we get to n eigenvectors is if every eigenvalue has a number of linearly independent eigenvectors equal to the algebraic multiplicity, which means they are not defective.

So how can we fix this? Well, there’s not really much we can do in the way of finding more eigenvectors, because they don’t exist. The replacement that we have is, in linear algebra contexts, called a *generalized eigenvector*. We will see this idea come back up in § 4.4 in a more natural way.

If r is an defective eigenvalue of the matrix A with eigenvector \vec{v} , a *generalized eigenvector* of A is a vector \vec{w} so that $(A - rI)\vec{w} = \vec{v}$. This is the same as the normal eigenvector equation with \vec{v} on the right-hand side instead of $\vec{0}$. Since $(A - rI)\vec{v} = \vec{0}$, this also means that

$$(A - rI)^2\vec{w} = 0.$$

More generally, a generalized eigenvector is a vector \vec{w} where there is a power $k \geq 1$ so that

$$(A - rI)^k\vec{w} = 0 \quad \text{but} \quad (A - rI)^{k-1}\vec{w} \neq 0.$$

It might seems strange where this comes from, but we will see why this formula makes more sense once we try to solve differential equations using matrices in § 4.4.

Example 3.6.7: Find a generalized eigenvector of eigenvalue 3 for the matrix

$$B = \begin{bmatrix} 4 & -1 \\ 1 & 2 \end{bmatrix}.$$

Solution: Previously, we found that $\vec{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is an eigenvector for B with eigenvalue 3. To find a generalized eigenvector, we need a vector \vec{w} so that

$$(B - 3I)\vec{w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Plugging in the matrix for $B - 3I$ gives that we need

$$\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Both of the rows of this matrix becomes the equation

$$w_1 - w_2 = 1.$$

There are many values of w_1 and w_2 that make this work. We can pick $w_1 = 1$ and $w_2 = 0$. This will give a generalized eigenvector of $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$. We could also pick $w_1 = 3$ and $w_2 = 2$, to get a generalized eigenvector as $\begin{bmatrix} 3 \\ 2 \end{bmatrix}$. Any of these choices work as a generalized eigenvector. \square

Example 3.6.8: Find the eigenvalues and eigenvectors (and generalized eigenvectors if needed) of the matrix

$$A = \begin{bmatrix} -2 & 0 & 1 \\ 19 & 2 & -16 \\ -1 & 0 & 0 \end{bmatrix}.$$

Solution: We start by looking for the eigenvalues through the characteristic polynomial.

$$\det(A - \lambda I) = \det \left(\begin{bmatrix} -2 - \lambda & 0 & 1 \\ 19 & 2 - \lambda & -16 \\ -1 & 0 & 0 - \lambda \end{bmatrix} \right)$$

To compute this determinant, we will expand along column 2, because it only has one non-zero entry. This gives

$$\begin{aligned} \det(A - \lambda I) &= (-1)^{2+2}(2 - \lambda) \det \left(\begin{bmatrix} -2 - \lambda & 1 \\ -1 & -\lambda \end{bmatrix} \right) \\ &= (2 - \lambda)((-2 - \lambda)(-\lambda) + 1) \\ &= (2 - \lambda)(\lambda^2 + 2\lambda + 1) = (2 - \lambda)(\lambda + 1)^2 \end{aligned}$$

so we have an eigenvalue at 2 and a double eigenvalue at -1 .

First, let's look for the eigenvector for eigenvalue 2. In this case, we know that the eigenvector must satisfy

$$\begin{bmatrix} -4 & 0 & 1 \\ 19 & 0 & -16 \\ -1 & 0 & -2 \end{bmatrix} \vec{v} = \vec{0}.$$

Row reducing the coefficient matrix will give

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

so that a corresponding eigenvector is

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

since we know that $v_1 = 0$ and $v_3 = 0$.

For $\lambda = -1$, we see that an eigenvector must satisfy

$$\begin{bmatrix} -1 & 0 & 1 \\ 19 & 3 & -16 \\ -1 & 0 & 1 \end{bmatrix} \vec{v} = \vec{0}.$$

We now look to row reduce this coefficient matrix.

$$\begin{aligned} \begin{bmatrix} -1 & 0 & 1 \\ 19 & 3 & -16 \\ -1 & 0 & 1 \end{bmatrix} &\rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 19 & 3 & -16 \\ -1 & 0 & 1 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 3 & 3 \\ 0 & 0 & 0 \end{bmatrix} . \\ &\rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

Therefore, we know that

$$v_1 - v_3 = 0 \quad v_2 + v_3 = 0.$$

If we pick $v_3 = 1$, then we know that $v_2 = -1$ and $v_1 = 1$, so the only eigenvector we get for $\lambda = -1$ is

$$\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} .$$

Since we only found one eigenvector for $\lambda = -1$ and $\lambda + 1$ was squared in the characteristic polynomial, this is a defective eigenvalue. Thus, we can look for a generalized eigenvalue here, which means that we need to solve for a vector \vec{w} with

$$\begin{bmatrix} -1 & 0 & 1 \\ 19 & 3 & -16 \\ -1 & 0 & 1 \end{bmatrix} \vec{w} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$$

We can then row reduce the augmented matrix to see what we can pick for \vec{w} .

$$\begin{bmatrix} -1 & 0 & 1 & 1 \\ 19 & 3 & -16 & -1 \\ -1 & 0 & 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & -1 & -1 \\ 19 & 3 & -16 & -1 \\ -1 & 0 & 1 & 1 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} 1 & 0 & -1 & -1 \\ 0 & 3 & 3 & 18 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} 1 & 0 & -1 & -1 \\ 0 & 1 & 1 & 6 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Thus, the generalized eigenvector \vec{w} must satisfy

$$w_1 - w_3 = -1 \quad w_2 + w_3 = 6.$$

We can pick any non-zero numbers to do this, so we can take $w_3 = 1$, $w_2 = 5$ and $w_1 = 0$. Thus, the generalized eigenvector here is

$$\begin{bmatrix} 0 \\ 5 \\ 1 \end{bmatrix}.$$

□

3.6.5 Exercises

Exercise 3.6.1:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} -8 & -18 \\ 4 & 10 \end{bmatrix}$$

Exercise 3.6.2:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} -2 & 0 \\ 8 & -4 \end{bmatrix}$$

Exercise 3.6.3:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} -7 & 1 \\ -12 & 0 \end{bmatrix}$$

Exercise 3.6.4:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} -3 & 5 \\ -8 & 9 \end{bmatrix}$$

Exercise 3.6.5:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} 0 & 2 \\ -1 & -2 \end{bmatrix}$$

Exercise 3.6.6:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} -4 & 1 \\ -8 & 0 \end{bmatrix}$$

Exercise 3.6.7:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} 0 & -8 \\ 2 & 8 \end{bmatrix}$$

Exercise 3.6.8:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} 1 & -2 \\ 8 & -7 \end{bmatrix}$$

Exercise 3.6.9:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} 4 & 0 & 0 \\ -4 & 2 & 1 \\ -6 & 0 & 1 \end{bmatrix}$$

Exercise 3.6.10:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} -4 & 9 & 9 \\ -3 & 6 & 9 \\ 3 & -7 & -10 \end{bmatrix}$$

Exercise 3.6.11:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} -2 & 0 & 0 \\ 0 & 4 & 6 \\ 6 & -3 & -2 \end{bmatrix}$$

Exercise 3.6.12:* Find the eigenvalues and corresponding eigenvectors of the matrix

$$\begin{bmatrix} 5 & 3 & 6 \\ 2 & 2 & 2 \\ -3 & -2 & -3 \end{bmatrix}$$

Exercise 3.6.13:* Find the eigenvalues and eigenvectors for the matrix below. Compute generalized eigenvectors if needed to get to a total of two vectors.

$$\begin{bmatrix} -11 & -9 \\ 12 & 10 \end{bmatrix}$$

Exercise 3.6.14:* Find the eigenvalues and eigenvectors for the matrix below. Compute generalized eigenvectors if needed to get to a total of two vectors.

$$\begin{bmatrix} 4 & -4 \\ 1 & 0 \end{bmatrix}$$

Exercise 3.6.15: We say that a matrix A is diagonalizable if there exist matrices D and P so that $PDP^{-1} = A$. This really means that A can be represented by a diagonal matrix in a different basis (as opposed to the standard basis). One way this can be done is with eigenvalues.

- a) Consider the matrix A given by

$$A = \begin{bmatrix} -4 & 6 \\ -1 & 1 \end{bmatrix}.$$

Find the eigenvalues and corresponding eigenvectors of this matrix.

- b) Form two matrices, D , a diagonal matrix with the eigenvalues of A on the diagonal, and E , a matrix whose columns are the eigenvectors of A in the same order as the eigenvalues were put into D . Write out these matrices.
- c) Compute E^{-1} .
- d) Work out the products EDE^{-1} and $E^{-1}AE$. What do you notice here?

This shows that, in the case of a 2×2 matrix, if we have two distinct real eigenvalues, that matrix is diagonalizable, using the eigenvectors.

Exercise 3.6.16: Follow the process outlined in Exercise 3.6.15 to attempt to diagonalize the matrix

$$\begin{bmatrix} 13 & 14 & 12 \\ -6 & -4 & -6 \\ -3 & -6 & -2 \end{bmatrix}$$

Hint: 1 is an eigenvalue.

Exercise 3.6.17: The diagonalization process described in Exercise 3.6.15 works for any case where there are real and distinct eigenvalues, as well as complex eigenvalues (but the algebra with the complex numbers gets complicated). It may or may not work in the case of repeated eigenvalues, and it fails whenever there are defective eigenvalues. Consider the matrix

$$\begin{bmatrix} 4 & -1 \\ 1 & 2 \end{bmatrix}$$

- a) Find the eigenvalue(s) of this matrix, and see that we have a repeated eigenvalue.
- b) Find the eigenvector for that eigenvalue, as well as a generalized eigenvector.
- c) Build a matrix E like before, but this time put the eigenvector in the first column and the generalized eigenvector in the second. Compute E^{-1} .
- d) Find the product $E^{-1}AE$. Before, this gave us a diagonal matrix, but what do we get now?

The matrix we get here is almost diagonal, but not quite. It turns out that this is the best we can do for matrices with defective eigenvalues. This matrix is often called J and is the Jordan Form of the matrix A .

Exercise 3.6.18:* Follow the process in Exercise 3.6.17 to find the Jordan Form of the matrix

$$\begin{bmatrix} -7 & 5 & 5 \\ -4 & 5 & 7 \\ -6 & 3 & 1 \end{bmatrix}.$$

3.7 Kernel and Nullity

Attribution: [JL], §A.4.

Learning Objectives

After this section, you will be able to:

- Determine the kernel of a matrix using row reduction and
- Understand the connection between rank and nullity in a given matrix.

3.7.1 Kernel

The set of solutions of a linear equation $L\vec{x} = \vec{0}$, the kernel of L , is a subspace: If \vec{x} and \vec{y} are solutions, then

$$L(\vec{x} + \vec{y}) = L\vec{x} + L\vec{y} = \vec{0} + \vec{0} = \vec{0}, \quad \text{and} \quad L(\alpha\vec{x}) = \alpha L\vec{x} = \alpha\vec{0} = \vec{0}.$$

So $\vec{x} + \vec{y}$ and $\alpha\vec{x}$ are solutions. The dimension of the kernel is called the *nullity* of the matrix.

The same sort of idea governs the solutions of linear differential equations. We try to describe the kernel of a linear differential operator, and as it is a subspace, we look for a basis of this kernel. Much of this book is dedicated to finding such bases.

The kernel of a matrix is the same as the kernel of its reduced row echelon form. For a matrix in reduced row echelon form, the kernel is rather easy to find. If a vector \vec{x} is applied to a matrix L , then each entry in \vec{x} corresponds to a column of L , the column that the entry multiplies. To find the kernel, pick a non-pivot column make a vector that has a -1 in the entry corresponding to this non-pivot column and zeros at all the other entries corresponding to the other non-pivot columns. Then for all the entries corresponding to pivot columns make it precisely the value in the corresponding row of the non-pivot column to make the vector be a solution to $L\vec{x} = \vec{0}$. This procedure is best understood by example.

Example 3.7.1: Consider

$$L = \begin{bmatrix} \boxed{1} & 2 & 0 & 0 & 3 \\ 0 & 0 & \boxed{1} & 0 & 4 \\ 0 & 0 & 0 & \boxed{1} & 5 \end{bmatrix}.$$

This matrix is in reduced row echelon form, the pivots are marked. There are two non-pivot columns, so the kernel has dimension 2, that is, it is the span of 2 vectors. Let us find the first vector. We look at the first non-pivot column, the 2nd column, and we put a -1 in the 2nd entry of our vector. We put a 0 in the 5th entry as the 5th column is also a non-pivot column:

$$\begin{bmatrix} ? \\ -1 \\ ? \\ ? \\ 0 \end{bmatrix}.$$

Let us fill the rest. When this vector hits the first row, we get a -2 and 1 times whatever the first question mark is. So make the first question mark 2. For the second and third rows, it is sufficient to make it the question marks zero. We are really filling in the non-pivot column into the remaining entries. Let us check while marking which numbers went where:

$$\begin{bmatrix} 1 & \boxed{2} & 0 & 0 & 3 \\ 0 & \boxed{0} & 1 & 0 & 4 \\ 0 & \boxed{0} & 0 & 1 & 5 \end{bmatrix} \begin{bmatrix} \boxed{2} \\ -1 \\ \boxed{0} \\ \boxed{0} \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Yay! How about the second vector. We start with

$$\begin{bmatrix} ? \\ 0 \\ ? \\ ? \\ -1 \end{bmatrix}$$

We set the first question mark to 3, the second to 4, and the third to 5. Let us check, marking things as previously,

$$\begin{bmatrix} 1 & 2 & 0 & 0 & \boxed{3} \\ 0 & 0 & 1 & 0 & \boxed{4} \\ 0 & 0 & 0 & 1 & \boxed{5} \end{bmatrix} \begin{bmatrix} \boxed{3} \\ 0 \\ \boxed{4} \\ \boxed{5} \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

There are two non-pivot columns, so we only need two vectors. We have found the basis of the kernel. So,

$$\text{kernel of } L = \text{span} \left\{ \begin{bmatrix} 2 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \\ 4 \\ 5 \\ -1 \end{bmatrix} \right\}$$

What we did in finding a basis of the kernel is we expressed all solutions of $L\vec{x} = \vec{0}$ as a linear combination of some given vectors.

The procedure to find the basis of the kernel of a matrix L :

- (i) Find the reduced row echelon form of L .
- (ii) Write down the basis of the kernel as above, one vector for each non-pivot column.

The rank of a matrix is the dimension of the column space, and that is the span on the pivot columns, while the kernel is the span of vectors one for each non-pivot column. So the two numbers must add to the number of columns.

Theorem 3.7.1 (Rank–Nullity)

If a matrix A has n columns, rank r , and nullity k (dimension of the kernel), then

$$n = r + k.$$

The theorem is immensely useful in applications. It allows one to compute the rank r if one knows the nullity k and vice versa, without doing any extra work.

Let us consider an example application, a simple version of the so-called *Fredholm alternative*. A similar result is true for differential equations. Consider

$$A\vec{x} = \vec{b},$$

where A is a square $n \times n$ matrix. There are then two mutually exclusive possibilities:

- (i) A nonzero solution \vec{x} to $A\vec{x} = \vec{0}$ exists.
- (ii) The equation $A\vec{x} = \vec{b}$ has a unique solution \vec{x} for every \vec{b} .

How does the Rank–Nullity theorem come into the picture? Well, if A has a nonzero solution \vec{x} to $A\vec{x} = \vec{0}$, then the nullity k is positive. But then the rank $r = n - k$ must be less than n . In particular it means that the column space of A is of dimension less than n , so it is a subspace that does not include everything in \mathbb{R}^n . So \mathbb{R}^n has to contain some vector \vec{b} not in the column space of A . In fact, most vectors in \mathbb{R}^n are not in the column space of A .

3.7.2 Connection with Eigenvalues and Eigenvectors

The idea of a kernel also comes up when defining and discussing eigenvectors. In order to find this vector, we are looking for a vector \vec{v} so that

$$(A - \lambda I)\vec{v} = \vec{0}.$$

This means that we are looking for a vector \vec{v} that is in the kernel of the matrix $(A - \lambda I)$. Since the kernel is also a subspace, this means that the set of all eigenvectors of a matrix A with a certain eigenvalue is a subspace, so it has a dimension. This dimension is number of linearly independent eigenvectors with that eigenvalue, so it is the geometric multiplicity of this eigenvalue. This also motivates why this is sometimes called the *eigenspace* for a given eigenvalue. Finding a basis of this subspace (which is also finding the kernel of the matrix $A - \lambda I$) is the exact same as the process of finding the eigenvectors of the matrix A .

3.7.3 Exercises

Exercise 3.7.1: For the following matrices, find a basis for the kernel (nullspace).

a) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 5 \\ 1 & 1 & -4 \end{bmatrix}$	b) $\begin{bmatrix} 2 & -1 & -3 \\ 4 & 0 & -4 \\ -1 & 1 & 2 \end{bmatrix}$	c) $\begin{bmatrix} -4 & 4 & 4 \\ -1 & 1 & 1 \\ -5 & 5 & 5 \end{bmatrix}$	d) $\begin{bmatrix} -2 & 1 & 1 & 1 \\ -4 & 2 & 2 & 2 \\ 1 & 0 & 4 & 3 \end{bmatrix}$
---	--	---	--

Exercise 3.7.2:* For the following matrices, find a basis for the kernel (nullspace).

$$\begin{array}{ll} \text{a) } \begin{bmatrix} 2 & 6 & 1 & 9 \\ 1 & 3 & 2 & 9 \\ 3 & 9 & 0 & 9 \end{bmatrix} & \text{b) } \begin{bmatrix} 2 & -2 & -5 \\ -1 & 1 & 5 \\ -5 & 5 & -3 \end{bmatrix} \quad \text{c) } \begin{bmatrix} 1 & -5 & -4 \\ 2 & 3 & 5 \\ -3 & 5 & 2 \end{bmatrix} \quad \text{d) } \begin{bmatrix} 0 & 4 & 4 \\ 0 & 1 & 1 \\ 0 & 5 & 5 \end{bmatrix} \end{array}$$

Exercise 3.7.3: Suppose a 5×5 matrix A has rank 3. What is the nullity?

Exercise 3.7.4: Consider a square matrix A , and suppose that \vec{x} is a nonzero vector such that $A\vec{x} = \vec{0}$. What does the Fredholm alternative say about invertibility of A .

Exercise 3.7.5: Consider

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 2 & ? & ? \\ -1 & ? & ? \end{bmatrix}.$$

If the nullity of this matrix is 2, fill in the question marks. Hint: What is the rank?

Exercise 3.7.6:* Suppose the column space of a 9×5 matrix A of dimension 3. Find

- a) Rank of A .
- b) Nullity of A .
- c) Dimension of the row space of A .
- d) Dimension of the nullspace of A .
- e) Size of the maximum subset of linearly independent rows of A .

Chapter 4

Systems of ODEs

4.1 Introduction to systems of ODEs

Attribution: [JL], §3.1.

Learning Objectives

After this section, you will be able to:

- Classify the order and number of components in a system of differential equations,
- Verify if a set of functions solves a system of differential equations, and
- Write a system of differential equations to fit a physical situation.

4.1.1 Systems

Often we do not have just one dependent variable and one equation. And as we will see, we may end up with systems of several equations and several dependent variables even if we start with a single equation.

If we have several dependent variables, suppose y_1, y_2, \dots, y_n , then we can have a differential equation involving all of them and their derivatives with respect to one independent variable x . For example, $y_1'' = f(y_1', y_2', y_1, y_2, x)$. Usually, when we have two dependent variables we have two equations such as

$$\begin{aligned}y_1'' &= f_1(y_1', y_2', y_1, y_2, x), \\y_2'' &= f_2(y_1', y_2', y_1, y_2, x),\end{aligned}$$

for some functions f_1 and f_2 . We call the above a *system of differential equations*. More precisely, the above is a *second order system* of ODEs as second order derivatives appear. The system

$$\begin{aligned}x_1' &= g_1(x_1, x_2, x_3, t), \\x_2' &= g_2(x_1, x_2, x_3, t), \\x_3' &= g_3(x_1, x_2, x_3, t),\end{aligned}$$

is a *first order system*, where x_1, x_2, x_3 are the dependent variables, and t is the independent variable.

The terminology for systems is essentially the same as for single equations. For the system above, a *solution* is a set of three functions $x_1(t), x_2(t), x_3(t)$, such that

$$\begin{aligned}x'_1(t) &= g_1(x_1(t), x_2(t), x_3(t), t), \\x'_2(t) &= g_2(x_1(t), x_2(t), x_3(t), t), \\x'_3(t) &= g_3(x_1(t), x_2(t), x_3(t), t).\end{aligned}$$

In order to verify that something is a solution, we plug the different components into the solution to see that all of the equations are satisfied; if any one of the equations is not satisfied, then this set of functions is not a solution. We usually also have an *initial condition*. Just like for single equations we specify x_1, x_2 , and x_3 for some fixed t . For example, $x_1(0) = a_1$, $x_2(0) = a_2$, $x_3(0) = a_3$. For some constants a_1, a_2 , and a_3 . For the second order system we would also specify the first derivatives at a point. And if we find a solution with constants in it, where by solving for the constants we find a solution for any initial condition, we call this solution the *general solution*. Best to look at a simple example.

Example 4.1.1: Sometimes a system is easy to solve by solving for one variable and then for the second variable. Take the first order system

$$\begin{aligned}y'_1 &= y_1, \\y'_2 &= y_1 - y_2,\end{aligned}$$

with y_1, y_2 as the dependent variables and x as the independent variable. Consider initial conditions $y_1(0) = 1, y_2(0) = 2$ and solve the initial value problem.

Solution: We note that $y_1 = C_1 e^x$ is the general solution of the first equation. We then plug this y_1 into the second equation and get the equation $y'_2 = C_1 e^x - y_2$, which is a linear first order equation that is easily solved for y_2 . By the method of integrating factor we get

$$e^x y_2 = \frac{C_1}{2} e^{2x} + C_2,$$

or $y_2 = \frac{C_1}{2} e^x + C_2 e^{-x}$. The general solution to the system is, therefore,

$$y_1 = C_1 e^x, \quad y_2 = \frac{C_1}{2} e^x + C_2 e^{-x}.$$

We solve for C_1 and C_2 given the initial conditions. We substitute $x = 0$ and find that $C_1 = 1$ and $C_2 = 3/2$. Thus the solution is $y_1 = e^x$, and $y_2 = (1/2)e^x + (3/2)e^{-x}$. \square

Generally, we will not be so lucky to be able to solve for each variable separately as in the example above, and we will have to solve for all variables at once. While we won't generally be able to solve for one variable and then the next, we will try to salvage as much as possible from this technique. It will turn out that in a certain sense we will still (try to) solve a bunch of single equations and put their solutions together. Let's not worry right now about how to solve systems yet.

We will mostly consider the *linear systems*. The example above is an example of a *linear first order system*. It is linear as none of the dependent variables or their derivatives appear in nonlinear functions or with powers higher than one (x , y , x' and y' , constants, and functions of t can appear, but not xy or $(y')^2$ or x^3). Another, more complicated, example of a linear system is

$$\begin{aligned} y_1'' &= e^t y_1' + t^2 y_1 + 5y_2 + \sin(t), \\ y_2'' &= t y_1' - y_2' + 2y_1 + \cos(t). \end{aligned}$$

4.1.2 Applications

Let us consider some simple applications of systems and how to set up the equations.

Example 4.1.2: First, we consider salt and brine tanks, but this time water flows from one to the other and back. We again consider that the tanks are evenly mixed.

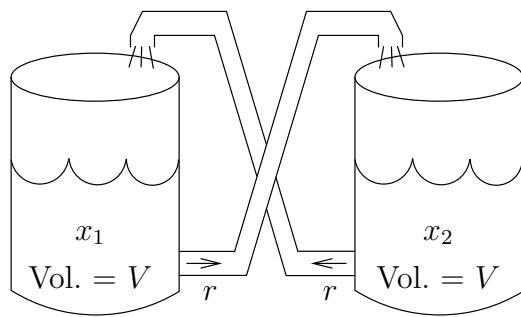


Figure 4.1: A closed system of two brine tanks.

Suppose we have two tanks, each containing volume V liters of salt brine. The amount of salt in the first tank is x_1 grams, and the amount of salt in the second tank is x_2 grams. The liquid is perfectly mixed and flows at the rate r liters per second out of each tank into the other. See Figure 4.1.

Solution: The rate of change of x_1 , that is x'_1 , is the rate of salt coming in minus the rate going out. The rate coming in is the density of the salt in tank 2, that is $\frac{x_2}{V}$, times the rate r . The rate coming out is the density of the salt in tank 1, that is $\frac{x_1}{V}$, times the rate r . In other words it is

$$x'_1 = \frac{x_2}{V}r - \frac{x_1}{V}r = \frac{r}{V}x_2 - \frac{r}{V}x_1 = \frac{r}{V}(x_2 - x_1).$$

Similarly we find the rate x'_2 , where the roles of x_1 and x_2 are reversed. All in all, the system of ODEs for this problem is

$$\begin{aligned} x'_1 &= \frac{r}{V}(x_2 - x_1), \\ x'_2 &= \frac{r}{V}(x_1 - x_2). \end{aligned}$$

In this system we cannot solve for x_1 or x_2 separately. We must solve for both x_1 and x_2 at once, which is intuitively clear since the amount of salt in one tank affects the amount in the other. We can't know x_1 before we know x_2 , and vice versa.

We don't yet know how to find all the solutions, but intuitively we can at least find some solutions. Suppose we know that initially the tanks have the same amount of salt. That is, we have an initial condition such as $x_1(0) = x_2(0) = C$. Then clearly the amount of salt coming and out of each tank is the same, so the amounts are not changing. In other words, $x_1 = C$ and $x_2 = C$ (the constant functions) is a solution: $x'_1 = x'_2 = 0$, and $x_2 - x_1 = x_1 - x_2 = 0$, so the equations are satisfied.

Let us think about the setup a little bit more without solving it. Suppose the initial conditions are $x_1(0) = A$ and $x_2(0) = B$, for two different constants A and B . Since no salt is coming in or out of this closed system, the total amount of salt is constant. That is, $x_1 + x_2$ is constant, and so it equals $A + B$. Intuitively if A is bigger than B , then more salt will flow out of tank one than into it. Eventually, after a long time we would then expect the amount of salt in each tank to equalize. In other words, the solutions of both x_1 and x_2 should tend towards $\frac{A+B}{2}$. Once you know how to solve systems you will find out that this really is so. \square

Example 4.1.3: Another example that showcases how systems work is different ways that populations of animals can interact. There are two main interactions that we will consider. The first of these is of two “competing species.” The idea here is that there are two species that are trying to coexist in a given area. On their own (without the other species), each one would grow exponentially, but any interaction between the two species is negative for both of them, because they share the types of food and other resources that they need to survive and grow. This gives rise to a system of differential equations of the form

$$\begin{aligned}\frac{dx_1}{dt} &= ax_1 - bx_1x_2 \\ \frac{dx_2}{dt} &= cx_2 - dx_1x_2\end{aligned}.$$

In the system here, the coefficient a represents the growth rate of species 1 on its own, b represents the amount to which the competition for resources affects the growth rate of species 1, c represents the growth rate of species 2, and d represents the magnitude of how the competition affects the growth of species 2. This type of system can also be written to contain logistic growth terms for the two species, resulting in

$$\begin{aligned}\frac{dx_1}{dt} &= ax_1(K_1 - x_1) - bx_1x_2 \\ \frac{dx_2}{dt} &= cx_2(K_2 - x_2) - dx_1x_2\end{aligned}.$$

The other main population model to consider is a “predator-prey” interaction. The key components of this model are that the prey population will grow on its own and the interaction between the two populations is negative, because the presence of predator population will cause the prey population to decrease. On the other hand, the predator population will die off on its own (without a food source) but the interaction with the prey population causes

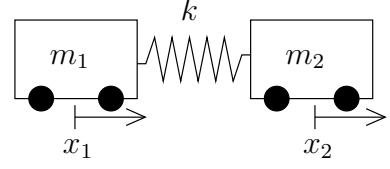
the predator population to increase. This gives rise to the system of differential equations

$$\begin{aligned}\frac{dx}{dt} &= ax - bxy \\ \frac{dy}{dt} &= -cy + dxy\end{aligned}$$

where x is the prey population and y is the predator population. We will take another look at both of these examples in § 5.3 once we have more terminology and techniques to discuss them.

Example 4.1.4: Let us look at a second order example. We return to the mass and spring setup, but this time we consider two masses.

Consider one spring with constant k and two masses m_1 and m_2 . Think of the masses as carts that ride along a straight track with no friction. Let x_1 be the displacement of the first cart and x_2 be the displacement of the second cart. That is, we put the two carts somewhere with no tension on the spring, and we mark the position of the first and second cart and call those the zero positions. Then x_1 measures how far the first cart is from its zero position, and x_2 measures how far the second cart is from its zero position. The force exerted by the spring on the first cart is $k(x_2 - x_1)$, since $x_2 - x_1$ is how far the string is stretched (or compressed) from the rest position. The force exerted on the second cart is the opposite, thus the same thing with a negative sign. Newton's second law states that force equals mass times acceleration. So the system of equations is



$$\begin{aligned}m_1 x_1'' &= k(x_2 - x_1), \\ m_2 x_2'' &= -k(x_2 - x_1).\end{aligned}$$

Again, we cannot solve for the x_1 or x_2 variable separately. That we must solve for both x_1 and x_2 at once is intuitively clear, since where the first cart goes depends on exactly where the second cart goes and vice versa.

4.1.3 Changing to first order

Before we talk about how to handle systems, let us note that in some sense we need only consider first order systems. Let us take an n^{th} order differential equation

$$y^{(n)} = F(y^{(n-1)}, \dots, y', y, x).$$

We define new variables u_1, u_2, \dots, u_n and write the system

$$\begin{aligned}u'_1 &= u_2, \\ u'_2 &= u_3, \\ &\vdots \\ u'_{n-1} &= u_n, \\ u'_n &= F(u_n, u_{n-1}, \dots, u_2, u_1, x).\end{aligned}$$

We solve this system for u_1, u_2, \dots, u_n . Once we have solved for the u 's, we can discard u_2 through u_n and let $y = u_1$. This y solves the original equation.

Example 4.1.5: Take $x''' = 2x'' + 8x' + x + t$. Letting $u_1 = x, u_2 = x', u_3 = x''$, we find the system:

$$u'_1 = u_2, \quad u'_2 = u_3, \quad u'_3 = 2u_3 + 8u_2 + u_1 + t.$$

A similar process can be followed for a system of higher order differential equations. For example, a system of k differential equations in k unknowns, all of order n , can be transformed into a first order system of $n \times k$ equations and $n \times k$ unknowns.

Example 4.1.6: Consider the system from the carts example,

$$m_1x''_1 = k(x_2 - x_1), \quad m_2x''_2 = -k(x_2 - x_1).$$

Let $u_1 = x_1, u_2 = x'_1, u_3 = x_2, u_4 = x'_2$. The second order system becomes the first order system

$$u'_1 = u_2, \quad m_1u'_2 = k(u_3 - u_1), \quad u'_3 = u_4, \quad m_2u'_4 = -k(u_3 - u_1).$$

Example 4.1.7: The idea works in reverse as well. Consider the system

$$x' = 2y - x, \quad y' = x,$$

where the independent variable is t . We wish to solve for the initial conditions $x(0) = 1, y(0) = 0$.

Solution: If we differentiate the second equation, we get $y'' = x'$. We know what x' is in terms of x and y , and we know that $x = y'$. So,

$$y'' = x' = 2y - x = 2y - y'.$$

We now have the equation $y'' + y' - 2y = 0$. We know how to solve this equation and we find that $y = C_1e^{-2t} + C_2e^t$. Once we have y , we use the equation $y' = x$ to get x .

$$x = y' = -2C_1e^{-2t} + C_2e^t.$$

We solve for the initial conditions $1 = x(0) = -2C_1 + C_2$ and $0 = y(0) = C_1 + C_2$. Hence, $C_1 = -C_2$ and $1 = 3C_2$. So $C_1 = -1/3$ and $C_2 = 1/3$. Our solution is

$$x = \frac{2e^{-2t} + e^t}{3}, \quad y = \frac{-e^{-2t} + e^t}{3}.$$

\(\square\)

Exercise 4.1.1: Plug in and check that this really is the solution.

It is useful to go back and forth between systems and higher order equations for other reasons. For example, software for solving ODE numerically (approximation) is generally for first order systems. So to use it, you have to take whatever ODE you want to solve and convert it to a first order system. In fact, it is not very hard to adapt computer code for the Euler or Runge–Kutta method for first order equations to handle first order systems. We essentially just treat the dependent variable not as a number but as a vector. In many mathematical computer languages there is almost no distinction in syntax.

4.1.4 Autonomous systems and vector fields

A system where the equations do not depend on the independent variable is called an *autonomous system*. For example the system $y' = 2y - x$, $y' = x$ is autonomous as t is the independent variable but does not appear in the equations.

For autonomous systems we can draw the so-called *direction field* or *vector field*, a plot similar to a slope field, but instead of giving a slope at each point, we give a direction (and a magnitude). The previous example, $x' = 2y - x$, $y' = x$, says that at the point (x, y) the direction in which we should travel to satisfy the equations should be the direction of the vector $(2y - x, x)$ with the speed equal to the magnitude of this vector. So we draw the vector $(2y - x, x)$ at the point (x, y) and we do this for many points on the xy -plane. For example, at the point $(1, 2)$ we draw the vector $(2(2) - 1, 1) = (3, 1)$, a vector pointing to the right and a little bit up, while at the point $(2, 1)$ we draw the vector $(2(1) - 2, 2) = (0, 2)$ a vector that points straight up. When drawing the vectors, we will scale down their size to fit many of them on the same direction field. We are mostly interested in their direction and relative size. See [Figure 4.2](#).

We can draw a path of the solution in the plane. Suppose the solution is given by $x = f(t)$, $y = g(t)$. We pick an interval of t (say $0 \leq t \leq 2$ for our example) and plot all the points $(f(t), g(t))$ for t in the selected range. The resulting picture is called the *phase portrait* (or phase plane portrait). The particular curve obtained is called the *trajectory* or *solution curve*. See an example plot in [Figure 4.3](#). In the figure the solution starts at $(1, 0)$ and travels along the vector field for a distance of 2 units of t . We solved this system precisely, so we compute $x(2)$ and $y(2)$ to find $x(2) \approx 2.475$ and $y(2) \approx 2.457$. This point corresponds to the top right end of the plotted solution curve in the figure.

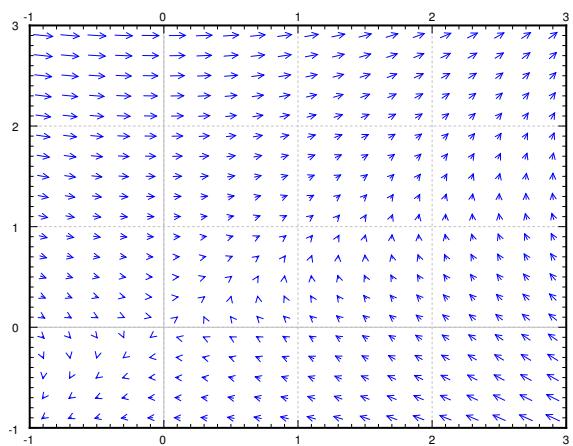


Figure 4.2: The direction field for $x' = 2y - x$, $y' = x$.

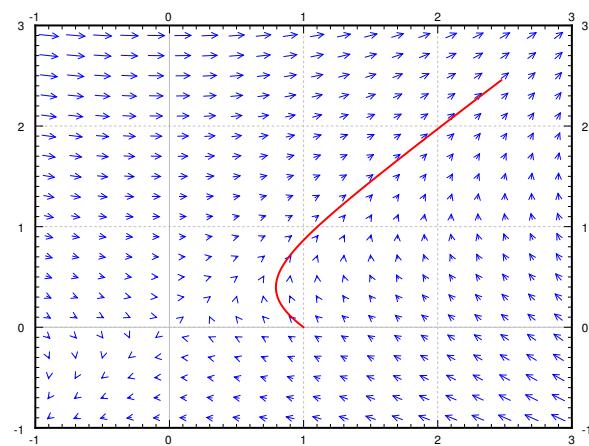


Figure 4.3: The direction field for $x' = 2y - x$, $y' = x$ with the trajectory of the solution starting at $(1, 0)$ for $0 \leq t \leq 2$.

Notice the similarity to the diagrams we drew for autonomous systems in one dimension. But note how much more complicated things become when we allow just one extra dimension.

We can draw phase portraits and trajectories in the xy -plane even if the system is not autonomous. In this case however we cannot draw the direction field, since the field changes as t changes. For each t we would get a different direction field.

4.1.5 Picard's theorem

Perhaps before going further, let us mention that Picard's theorem on existence and uniqueness still holds for systems of ODE. Let us restate this theorem in the setting of systems. A general first order system is of the form

$$\begin{aligned} x'_1 &= F_1(x_1, x_2, \dots, x_n, t), \\ x'_2 &= F_2(x_1, x_2, \dots, x_n, t), \\ &\vdots \\ x'_n &= F_n(x_1, x_2, \dots, x_n, t). \end{aligned} \tag{4.1}$$

Theorem 4.1.1 (Picard's theorem on existence and uniqueness for systems)

If for every $j = 1, 2, \dots, n$ and every $k = 1, 2, \dots, n$ each F_j is continuous and the derivative $\frac{\partial F_j}{\partial x_k}$ exists and is continuous near some $(x_1^0, x_2^0, \dots, x_n^0, t^0)$, then a solution to (4.1) subject to the initial condition $x_1(t^0) = x_1^0, x_2(t^0) = x_2^0, \dots, x_n(t^0) = x_n^0$ exists (at least for some small interval of t 's) and is unique.

That is, a unique solution exists for any initial condition given that the system is reasonable (F_j and its partial derivatives in the x variables are continuous). As for single equations we may not have a solution for all time t , but at least for some short period of time.

As we can change any n th order ODE into a first order system, then we notice that this theorem provides also the existence and uniqueness of solutions for higher order equations that we have until now not stated explicitly.

4.1.6 Exercises

Exercise 4.1.2: Verify that $x_1(t) = 2e^{-t} - 2e^{-2t}, x_2(t) = e^{-t} - 2e^{-2t}$ solves the system $x'_1 = -2x_2, x'_2 = x_1 - 3x_2$.

Exercise 4.1.3: Verify that $x_1(t) = -2te^{-3t} - 2e^{-3t}, x_2(t) = 2te^{-3t} + 3e^{-3t}$ solves the system $x'_1 = -5x_1 - 2x_2, x'_2 = 2x_1 - x_2$.

Exercise 4.1.4: Find the general solution of $x'_1 = x_2 - x_1 + t, x'_2 = x_2$.

Exercise 4.1.5: Find the general solution of $x'_1 = 3x_1 - x_2 + e^t, x'_2 = x_1$.

Exercise 4.1.6:* Find the general solution to $y'_1 = 3y_1, y'_2 = y_1 + y_2, y'_3 = y_1 + y_3$.

Exercise 4.1.7:* Solve $y' = 2x, x' = x + y, x(0) = 1, y(0) = 3$.

Exercise 4.1.8: Write $ay'' + by' + cy = f(x)$ as a first order system of ODEs.

Exercise 4.1.9: Write $x'' + y^2y' - x^3 = \sin(t)$, $y'' + (x' + y')^2 - x = 0$ as a first order system of ODEs.

Exercise 4.1.10:* Write $x''' = x + t$ as a first order system.

Exercise 4.1.11:* Write $y_1'' + y_1 + y_2 = t$, $y_2'' + y_1 - y_2 = t^2$ as a first order system.

Exercise 4.1.12: Write $y^{(4)} - t^2y''' + e^t y' - (2t + 1)y = \cos(t)$ as a first order system.

Exercise 4.1.13: Write the initial value problem

$$y'' - 2xy' + 3y = \sin(x) \quad y(0) = 1, \quad y'(0) = -2$$

as an initial value problem for a first order system of ODEs. Make sure to indicate how the initial condition appears as a part of this problem.

Exercise 4.1.14: Write the initial value problem

$$y'' - 2xy' + 3y = \sin(x) \quad y(0) = 1, \quad y'(0) = -2$$

as an initial value problem for a first order system of ODEs. Make sure to indicate how the initial condition appears as a part of this problem.

Exercise 4.1.15: Write the initial value problem

$$y^{(4)} + e^x y'' - 4 \cos(x) y' + (x^2 + 1)y = \frac{1}{x-3} \quad y(0) = 2, \quad y'(0) = -3, \quad y''(0) = 0, \quad y^{(3)}(0) = 1$$

as an initial value problem for a first order system of ODEs. Make sure to indicate how the initial condition appears as a part of this problem.

Exercise 4.1.16: Suppose two masses on carts on frictionless surface are at displacements x_1 and x_2 as in [Example 4.1.4](#) on page 241. Suppose that a rocket applies force F in the positive direction on cart x_1 . Set up the system of equations.

Exercise 4.1.17:* Suppose two masses on carts on frictionless surface are at displacements x_1 and x_2 as in [Example 4.1.4](#) on page 241. Suppose initial displacement is $x_1(0) = x_2(0) = 0$, and initial velocity is $x'_1(0) = x'_2(0) = a$ for some number a . Use your intuition to solve the system, explain your reasoning.

Exercise 4.1.18: Suppose the tanks are as in [Example 4.1.2](#) on page 239, starting both at volume V , but now the rate of flow from tank 1 to tank 2 is r_1 , and rate of flow from tank 2 to tank one is r_2 . In particular, the volumes will now be changing. Set up the system of equations.

Exercise 4.1.19:* Suppose the tanks are as in [Example 4.1.2](#) on page 239 except that clean water flows in at the rate s liters per second into tank 1, and brine flows out of tank 2 and into the sewer also at the rate of s liters per second.

- a) Draw the picture.
- b) Set up the system of equations.
- c) Intuitively, what happens as t goes to infinity, explain.

4.2 Matrices and linear systems

Attribution: [JL], §3.2.

Learning Objectives

After this section, you will be able to:

- Define and perform addition and multiplication operations on matrices,
- Compute the determinant of a square matrix, and
- Find eigenvalues and eigenvectors of a square matrix.

4.2.1 Matrices and vectors

Before we start talking about linear systems of ODEs, we need to talk about matrices, so let us review these briefly. A *matrix* is an $m \times n$ array of numbers (m rows and n columns). For example, we denote a 3×5 matrix as follows

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \end{bmatrix}.$$

The numbers a_{ij} are called *elements* or *entries*.

By a *vector* we usually mean a *column vector*, that is an $m \times 1$ matrix. If we mean a *row vector*, we will explicitly say so (a row vector is a $1 \times n$ matrix). We usually denote matrices by upper case letters and vectors by lower case letters with an arrow such as \vec{x} or \vec{b} . By $\vec{0}$ we mean the vector of all zeros.

We define some operations on matrices. We want 1×1 matrices to really act like numbers, so our operations have to be compatible with this viewpoint.

First, we can multiply a matrix by a *scalar* (a number). We simply multiply each entry in the matrix by the scalar. For example,

$$2 \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{bmatrix}.$$

Matrix addition is also easy. We add matrices element by element. For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} + \begin{bmatrix} 1 & 1 & -1 \\ 0 & 2 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 3 & 2 \\ 4 & 7 & 10 \end{bmatrix}.$$

If the sizes do not match, then addition is not defined.

If we denote by 0 the matrix with all zero entries, by c, d scalars, and by A, B, C matrices,

we have the following familiar rules:

$$\begin{aligned} A + 0 &= A = 0 + A, \\ A + B &= B + A, \\ (A + B) + C &= A + (B + C), \\ c(A + B) &= cA + cB, \\ (c + d)A &= cA + dA. \end{aligned}$$

Another useful operation for matrices is the so-called *transpose*. This operation just swaps rows and columns of a matrix. The transpose of A is denoted by A^T . Example:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}^T = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

4.2.2 Matrix multiplication

Let us now define matrix multiplication. First we define the so-called *dot product* (or *inner product*) of two vectors. Usually this will be a row vector multiplied with a column vector of the same size. For the dot product we multiply each pair of entries from the first and the second vector and we sum these products. The result is a single number. For example,

$$\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3.$$

And similarly for larger (or smaller) vectors.

Armed with the dot product we define the *product of matrices*. First let us denote by $\text{row}_i(A)$ the i^{th} row of A and by $\text{column}_j(A)$ the j^{th} column of A . For an $m \times n$ matrix A and an $n \times p$ matrix B we can define the product AB . We let AB be an $m \times p$ matrix whose ij^{th} entry is the dot product

$$\text{row}_i(A) \cdot \text{column}_j(B).$$

Do note how the sizes match up: $m \times n$ multiplied by $n \times p$ is $m \times p$. Example:

$$\begin{aligned} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} &= \\ &= \begin{bmatrix} 1 \cdot 1 + 2 \cdot 1 + 3 \cdot 1 & 1 \cdot 0 + 2 \cdot 1 + 3 \cdot 0 & 1 \cdot (-1) + 2 \cdot 1 + 3 \cdot 0 \\ 4 \cdot 1 + 5 \cdot 1 + 6 \cdot 1 & 4 \cdot 0 + 5 \cdot 1 + 6 \cdot 0 & 4 \cdot (-1) + 5 \cdot 1 + 6 \cdot 0 \end{bmatrix} = \begin{bmatrix} 6 & 2 & 1 \\ 15 & 5 & 1 \end{bmatrix} \end{aligned}$$

For multiplication we want an analogue of a 1. This analogue is the so-called *identity matrix*. The identity matrix is a square matrix with 1s on the diagonal and zeros everywhere else. It is usually denoted by I . For each size we have a different identity matrix and so

sometimes we may denote the size as a subscript. For example, the I_3 would be the 3×3 identity matrix

$$I = I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

We have the following rules for matrix multiplication. Suppose that A, B, C are matrices of the correct sizes so that the following make sense. Let α denote a scalar (number).

$$\begin{aligned} A(BC) &= (AB)C, \\ A(B + C) &= AB + AC, \\ (B + C)A &= BA + CA, \\ \alpha(AB) &= (\alpha A)B = A(\alpha B), \\ IA &= A = AI. \end{aligned}$$

A few warnings are in order.

- (i) $AB \neq BA$ in general (it may be true by fluke sometimes). That is, matrices do not commute. For example, take $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$.
- (ii) $AB = AC$ does not necessarily imply $B = C$, even if A is not 0.
- (iii) $AB = 0$ does not necessarily mean that $A = 0$ or $B = 0$. Try, for example, $A = B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$.

For the last two items to hold we would need to “divide” by a matrix. This is where the *matrix inverse* comes in. Suppose that A and B are $n \times n$ matrices such that

$$AB = I = BA.$$

Then we call B the inverse of A and we denote B by A^{-1} . If the inverse of A exists, then we call A *invertible*. If A is not invertible, we sometimes say A is *singular*.

If A is invertible, then $AB = AC$ does imply that $B = C$ (in particular the inverse of A is unique). We just multiply both sides by A^{-1} (on the left) to get $A^{-1}AB = A^{-1}AC$ or $IB = IC$ or $B = C$. It is also not hard to see that $(A^{-1})^{-1} = A$.

4.2.3 The determinant

For square matrices we define a useful quantity called the *determinant*. We define the determinant of a 1×1 matrix as the value of its only entry. For a 2×2 matrix we define

$$\det \left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) \stackrel{\text{def}}{=} ad - bc.$$

Before trying to define the determinant for larger matrices, let us note the meaning of the determinant. Consider an $n \times n$ matrix as a mapping of the n -dimensional euclidean space \mathbb{R}^n to itself, where \vec{x} gets sent to $A\vec{x}$. In particular, a 2×2 matrix A is a mapping of the plane

to itself. The determinant of A is the factor by which the area of objects changes. If we take the unit square (square of side 1) in the plane, then A takes the square to a parallelogram of area $|\det(A)|$. The sign of $\det(A)$ denotes changing of orientation (negative if the axes get flipped). For example, let

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

Then $\det(A) = 1 + 1 = 2$. Let us see where the (unit) square with vertices $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$ gets sent. Clearly $(0, 0)$ gets sent to $(0, 0)$.

$$\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

The image of the square is another square with vertices $(0, 0)$, $(1, -1)$, $(1, 1)$, and $(2, 0)$. The image square has a side of length $\sqrt{2}$ and is therefore of area 2.

If you think back to high school geometry, you may have seen a formula for computing the area of a parallelogram with vertices $(0, 0)$, (a, c) , (b, d) and $(a+b, c+d)$. And it is precisely

$$\left| \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} \right|.$$

The vertical lines above mean absolute value. The matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ carries the unit square to the given parallelogram.

Let us look at the determinant for larger matrices. We define A_{ij} as the matrix A with the i^{th} row and the j^{th} column deleted. To compute the determinant of a matrix, pick one row, say the i^{th} row and compute:

$$\det(A) = \sum_{j=1}^n (-1)^{i+j} a_{ij} \det(A_{ij}).$$

For the first row we get

$$\det(A) = a_{11} \det(A_{11}) - a_{12} \det(A_{12}) + a_{13} \det(A_{13}) - \dots \begin{cases} +a_{1n} \det(A_{1n}) & \text{if } n \text{ is odd,} \\ -a_{1n} \det(A_{1n}) & \text{if } n \text{ even.} \end{cases}$$

We alternately add and subtract the determinants of the submatrices A_{ij} multiplied by a_{ij} for a fixed i and all j . For a 3×3 matrix, picking the first row, we get $\det(A) = a_{11} \det(A_{11}) - a_{12} \det(A_{12}) + a_{13} \det(A_{13})$. For example,

$$\begin{aligned} \det \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} &= 1 \cdot \det \begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix} - 2 \cdot \det \begin{pmatrix} 4 & 6 \\ 7 & 9 \end{pmatrix} + 3 \cdot \det \begin{pmatrix} 4 & 5 \\ 7 & 8 \end{pmatrix} \\ &= 1(5 \cdot 9 - 6 \cdot 8) - 2(4 \cdot 9 - 6 \cdot 7) + 3(4 \cdot 8 - 5 \cdot 7) = 0. \end{aligned}$$

The numbers $(-1)^{i+j} \det(A_{ij})$ are called *cofactors* of the matrix and this way of computing the determinant is called the *cofactor expansion*. No matter which row you pick, you always

get the same number. It is also possible to compute the determinant by expanding along columns (picking a column instead of a row above). It is true that $\det(A) = \det(A^T)$.

A common notation for the determinant is a pair of vertical lines:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = \det \left(\begin{bmatrix} a & b \\ c & d \end{bmatrix} \right).$$

I personally find this notation confusing as vertical lines usually mean a positive quantity, while determinants can be negative. Also think about how to write the absolute value of a determinant. I will not use this notation in this book.

Think of the determinants telling you the scaling of a mapping. If B doubles the sizes of geometric objects and A triples them, then AB (which applies B to an object and then A) should make size go up by a factor of 6. This is true in general:

$$\det(AB) = \det(A)\det(B).$$

This property is one of the most useful, and it is employed often to actually compute determinants. A particularly interesting consequence is to note what it means for existence of inverses. Take A and B to be inverses of each other, that is $AB = I$. Then

$$\det(A)\det(B) = \det(AB) = \det(I) = 1.$$

Neither $\det(A)$ nor $\det(B)$ can be zero. Let us state this as a theorem as it will be very important in the context of this course.

Theorem 4.2.1

An $n \times n$ matrix A is invertible if and only if $\det(A) \neq 0$.

In fact, $\det(A^{-1})\det(A) = 1$ says that $\det(A^{-1}) = \frac{1}{\det(A)}$. So we even know what the determinant of A^{-1} is before we know how to compute A^{-1} .

There is a simple formula for the inverse of a 2×2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Notice the determinant of the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ in the denominator of the fraction. The formula only works if the determinant is nonzero, otherwise we are dividing by zero.

4.2.4 Solving linear systems

One application of matrices we will need is to solve systems of linear equations. This is best shown by example. Suppose that we have the following system of linear equations

$$\begin{aligned} 2x_1 + 2x_2 + 2x_3 &= 2, \\ x_1 + x_2 + 3x_3 &= 5, \\ x_1 + 4x_2 + x_3 &= 10. \end{aligned}$$

Without changing the solution, we could swap equations in this system, we could multiply any of the equations by a nonzero number, and we could add a multiple of one equation to another equation. It turns out these operations always suffice to find a solution.

It is easier to write the system as a matrix equation. The system above can be written as

$$\begin{bmatrix} 2 & 2 & 2 \\ 1 & 1 & 3 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \\ 10 \end{bmatrix}.$$

To solve the system we put the coefficient matrix (the matrix on the left-hand side of the equation) together with the vector on the right and side and get the so-called *augmented matrix*

$$\left[\begin{array}{ccc|c} 2 & 2 & 2 & 2 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{array} \right].$$

We apply the following three elementary operations.

- (i) Swap two rows.
- (ii) Multiply a row by a nonzero number.
- (iii) Add a multiple of one row to another row.

We keep doing these operations until we get into a state where it is easy to read off the answer, or until we get into a contradiction indicating no solution, for example if we come up with an equation such as $0 = 1$.

Let us work through the example. First multiply the first row by $\frac{1}{2}$ to obtain

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & 5 \\ 1 & 4 & 1 & 10 \end{array} \right].$$

Now subtract the first row from the second and third row.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 4 \\ 0 & 3 & 0 & 9 \end{array} \right]$$

Multiply the last row by $\frac{1}{3}$ and the second row by $\frac{1}{2}$.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 3 \end{array} \right]$$

Swap rows 2 and 3.

$$\left[\begin{array}{ccc|c} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{array} \right]$$

Subtract the last row from the first, then subtract the second row from the first.

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & -4 \\ 0 & 1 & 0 & 3 \\ 0 & 0 & 1 & 2 \end{array} \right]$$

If we think about what equations this augmented matrix represents, we see that $x_1 = -4$, $x_2 = 3$, and $x_3 = 2$. We try this solution in the original system and, voilà, it works!

Exercise 4.2.1: Check that the solution above really solves the given equations.

We write this equation in matrix notation as

$$A\vec{x} = \vec{b},$$

where A is the matrix $\begin{bmatrix} 2 & 2 & 2 \\ 1 & 1 & 3 \\ 1 & 4 & 1 \end{bmatrix}$ and \vec{b} is the vector $\begin{bmatrix} 2 \\ 5 \\ 10 \end{bmatrix}$. The solution can also be computed via the inverse,

$$\vec{x} = A^{-1}A\vec{x} = A^{-1}\vec{b}.$$

It is possible that the solution is not unique, or that no solution exists. It is easy to tell if a solution does not exist. If during the row reduction you come up with a row where all the entries except the last one are zero (the last entry in a row corresponds to the right-hand side of the equation), then the system is *inconsistent* and has no solution. For example, for a system of 3 equations and 3 unknowns, if you find a row such as $[0 \ 0 \ 0 \ | \ 1]$ in the augmented matrix, you know the system is inconsistent. That row corresponds to $0 = 1$.

You generally try to use row operations until the following conditions are satisfied. The first (from the left) nonzero entry in each row is called the *leading entry*.

- (i) The leading entry in any row is strictly to the right of the leading entry of the row above.
- (ii) Any zero rows are below all the nonzero rows.
- (iii) All leading entries are 1.
- (iv) All the entries above and below a leading entry are zero.

Such a matrix is said to be in *reduced row echelon form*. The variables corresponding to columns with no leading entries are said to be *free variables*. Free variables mean that we can pick those variables to be anything we want and then solve for the rest of the unknowns.

Example 4.2.1: The following augmented matrix is in reduced row echelon form.

$$\left[\begin{array}{ccc|c} 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{array} \right]$$

Suppose the variables are x_1 , x_2 , and x_3 . Then x_2 is the free variable, $x_1 = 3 - 2x_2$, and $x_3 = 1$.

On the other hand if during the row reduction process you come up with the matrix

$$\left[\begin{array}{ccc|c} 1 & 2 & 13 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 3 \end{array} \right],$$

there is no need to go further. The last row corresponds to the equation $0x_1 + 0x_2 + 0x_3 = 3$, which is preposterous. Hence, no solution exists.

4.2.5 Computing the inverse

If the matrix A is square and there exists a unique solution \vec{x} to $A\vec{x} = \vec{b}$ for any \vec{b} (there are no free variables), then A is invertible. Multiplying both sides by A^{-1} , you can see that $\vec{x} = A^{-1}\vec{b}$. So it is useful to compute the inverse if you want to solve the equation for many different right-hand sides \vec{b} .

We have a formula for the 2×2 inverse, but it is also not hard to compute inverses of larger matrices. While we will not have too much occasion to compute inverses for larger matrices than 2×2 by hand, let us touch on how to do it. Finding the inverse of A is actually just solving a bunch of linear equations. If we can solve $A\vec{x}_k = \vec{e}_k$ where \vec{e}_k is the vector with all zeros except a 1 at the k^{th} position, then the inverse is the matrix with the columns \vec{x}_k for $k = 1, 2, \dots, n$ (exercise: why?). Therefore, to find the inverse we write a larger $n \times 2n$ augmented matrix $[A | I]$, where I is the identity matrix. We then perform row reduction. The reduced row echelon form of $[A | I]$ will be of the form $[I | A^{-1}]$ if and only if A is invertible. We then just read off the inverse A^{-1} .

4.2.6 Eigenvalues and eigenvectors of a matrix

Let A be a constant square matrix. Suppose there is a scalar λ and a nonzero vector \vec{v} such that

$$A\vec{v} = \lambda\vec{v}.$$

We call λ an *eigenvalue* of A and we call \vec{v} a corresponding *eigenvector*.

Example 4.2.2: The matrix $\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$ has an eigenvalue $\lambda = 2$ with a corresponding eigenvector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ as

$$\begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Let us see how to compute eigenvalues for any matrix. Rewrite the equation for an eigenvalue as

$$(A - \lambda I)\vec{v} = \vec{0}.$$

This equation has a nonzero solution \vec{v} only if $A - \lambda I$ is not invertible. Were it invertible, we could write $(A - \lambda I)^{-1}(A - \lambda I)\vec{v} = (A - \lambda I)^{-1}\vec{0}$, which implies $\vec{v} = \vec{0}$. Therefore, A has the eigenvalue λ if and only if λ solves the equation

$$\det(A - \lambda I) = 0.$$

Consequently, we will be able to find an eigenvalue of A without finding a corresponding eigenvector. An eigenvector will have to be found later, once λ is known.

Example 4.2.3: Find all eigenvalues of $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$.

Solution: We write

$$\det \left(\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) = \det \left(\begin{bmatrix} 2 - \lambda & 1 & 1 \\ 1 & 2 - \lambda & 0 \\ 0 & 0 & 2 - \lambda \end{bmatrix} \right) = (2 - \lambda)((2 - \lambda)^2 - 1) = -(\lambda - 1)(\lambda - 2)(\lambda - 3).$$

So the eigenvalues are $\lambda = 1$, $\lambda = 2$, and $\lambda = 3$. □

For an $n \times n$ matrix, the polynomial we get by computing $\det(A - \lambda I)$ is of degree n , and hence in general, we have n eigenvalues. Some may be repeated, some may be complex.

To find an eigenvector corresponding to an eigenvalue λ , we write

$$(A - \lambda I)\vec{v} = \vec{0},$$

and solve for a nontrivial (nonzero) vector \vec{v} . If λ is an eigenvalue, there will be at least one free variable, and so for each distinct eigenvalue λ , we can always find an eigenvector.

Example 4.2.4: Find an eigenvector of $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ corresponding to the eigenvalue $\lambda = 3$.

Solution: We write

$$(A - \lambda I)\vec{v} = \left(\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} - 3 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 1 \\ 1 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \vec{0}.$$

It is easy to solve this system of linear equations. We write down the augmented matrix

$$\left[\begin{array}{ccc|c} -1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{array} \right],$$

and perform row operations (exercise: which ones?) until we get:

$$\left[\begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right].$$

The entries of \vec{v} have to satisfy the equations $v_1 - v_2 = 0$, $v_3 = 0$, and v_2 is a free variable. We can pick v_2 to be arbitrary (but nonzero), let $v_1 = v_2$, and of course $v_3 = 0$. For example, if we pick $v_2 = 1$, then $\vec{v} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$. Let us verify that \vec{v} really is an eigenvector corresponding to $\lambda = 3$:

$$\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 0 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

Yay! It worked. □

Exercise 4.2.2 (easy): Are eigenvectors unique? Can you find a different eigenvector for $\lambda = 3$ in the example above? How are the two eigenvectors related?

Exercise 4.2.3: When the matrix is 2×2 you do not need to do row operations when computing an eigenvector, you can read it off from $A - \lambda I$ (if you have computed the eigenvalues correctly). Can you see why? Explain. Try it for the matrix $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.

4.2.7 Exercises

Exercise 4.2.4: Let A and B be the matrices below.

$$A = \begin{bmatrix} 1 & 4 & -1 \\ 2 & 0 & 3 \\ 1 & -2 & 3 \end{bmatrix} \quad B = \begin{bmatrix} 0 & 2 & 3 \\ 1 & -4 & -2 \\ 2 & -5 & 1 \end{bmatrix}$$

Compute $A + 3B$, AB , and BA .

Exercise 4.2.5: Solve $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \vec{x} = \begin{bmatrix} 5 \\ 6 \end{bmatrix}$ by using matrix inverse.

Exercise 4.2.6: Compute determinant of $\begin{bmatrix} 9 & -2 & -6 \\ -8 & 3 & 6 \\ 10 & -2 & -6 \end{bmatrix}$.

Exercise 4.2.7:* Compute determinant of $\begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & -5 \\ 1 & -1 & 0 \end{bmatrix}$

Exercise 4.2.8: Compute determinant of $\begin{bmatrix} 1 & 2 & 3 & 1 \\ 4 & 0 & 5 & 0 \\ 6 & 0 & 7 & 0 \\ 8 & 0 & 10 & 1 \end{bmatrix}$. Hint: Expand along the proper row or column to make the calculations simpler.

Exercise 4.2.9: Compute inverse of $\begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$.

Exercise 4.2.10: For which h is $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & h \end{bmatrix}$ not invertible? Is there only one such h ? Are there several? Infinitely many?

Exercise 4.2.11:* Find t such that $\begin{bmatrix} 1 & t \\ -1 & 2 \end{bmatrix}$ is not invertible.

Exercise 4.2.12: For which h is $\begin{bmatrix} h & 1 & 1 \\ 0 & h & 0 \\ 1 & 1 & h \end{bmatrix}$ not invertible? Find all such h .

Exercise 4.2.13: Solve $\begin{bmatrix} 9 & -2 & -6 \\ -8 & 3 & 6 \\ 10 & -2 & -6 \end{bmatrix} \vec{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$.

Exercise 4.2.14: Solve $\begin{bmatrix} 5 & 3 & 7 \\ 8 & 4 & 4 \\ 6 & 3 & 3 \end{bmatrix} \vec{x} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}$.

Exercise 4.2.15:* Solve $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \vec{x} = \begin{bmatrix} 10 \\ 20 \end{bmatrix}$.

Exercise 4.2.16: Solve $\begin{bmatrix} 3 & 2 & 3 & 0 \\ 3 & 3 & 3 & 3 \\ 0 & 2 & 4 & 2 \\ 2 & 3 & 4 & 3 \end{bmatrix} \vec{x} = \begin{bmatrix} 2 \\ 0 \\ 4 \\ 1 \end{bmatrix}$.

Exercise 4.2.17: Find 3 nonzero 2×2 matrices A , B , and C such that $AB = AC$ but $B \neq C$.

Exercise 4.2.18:* Suppose a, b, c are nonzero numbers. Let $M = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$, $N = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}$.

- a) Compute M^{-1} . b) Compute N^{-1} .

Exercise 4.2.19 (easy): Let A be a 3×3 matrix with an eigenvalue of 3 and a corresponding eigenvector $\vec{v} = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}$. Find $A\vec{v}$.

Exercise 4.2.20:* Find the eigenvalues and eigenvectors for the matrix

$$\begin{bmatrix} 0 & -2 \\ 1 & 3 \end{bmatrix}.$$

Exercise 4.2.21:* Find the eigenvalues and eigenvectors for the matrix

$$\begin{bmatrix} -8 & -5 \\ 8 & 4 \end{bmatrix}.$$

Exercise 4.2.22:* Find the eigenvalues and eigenvectors for the matrix

$$\begin{bmatrix} 7 & -3 & 7 \\ 9 & -5 & 7 \\ 0 & 0 & -3 \end{bmatrix}.$$

4.3 Linear systems of ODEs

Attribution: [JL], §3.3.

Learning Objectives

After this section, you will be able to:

- Use proper terminology when discussing linear systems of differential equations and their solutions,
- Determine whether a set of functions is linearly independent, and
- Understand how the theory of non-homogeneous linear systems relates to the theory of non-linear equations.

First let us talk about matrix- or vector-valued functions. Such a function is just a matrix or vector whose entries depend on some variable. If t is the independent variable, we write a *vector-valued function* $\vec{x}(t)$ as

$$\vec{x}(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix}.$$

Similarly a *matrix-valued function* $A(t)$ is

$$A(t) = \begin{bmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \cdots & a_{2n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(t) & a_{n2}(t) & \cdots & a_{nn}(t) \end{bmatrix}.$$

The derivative $A'(t)$ or $\frac{dA}{dt}$ is just the matrix-valued function whose ij^{th} entry is $a'_{ij}(t)$.

Rules of differentiation of matrix-valued functions are similar to rules for normal functions. Let $A(t)$ and $B(t)$ be matrix-valued functions. Let c a scalar and let C be a constant matrix. Then

$$\begin{aligned} (A(t) + B(t))' &= A'(t) + B'(t), \\ (A(t)B(t))' &= A'(t)B(t) + A(t)B'(t), \\ (cA(t))' &= cA'(t), \\ (CA(t))' &= CA'(t), \\ (A(t)C)' &= A'(t)C. \end{aligned}$$

Note the order of the multiplication in the last two expressions.

A *first order linear system of ODEs* is a system that can be written as the vector equation

$$\vec{x}'(t) = P(t)\vec{x}(t) + \vec{f}(t),$$

where $P(t)$ is a matrix-valued function, and $\vec{x}(t)$ and $\vec{f}(t)$ are vector-valued functions. We will often suppress the dependence on t and only write $\vec{x}' = P\vec{x} + \vec{f}$. A solution of the system is a vector-valued function \vec{x} satisfying the vector equation.

For example, the equations

$$\begin{aligned}x'_1 &= 2tx_1 + e^t x_2 + t^2, \\x'_2 &= \frac{x_1}{t} - x_2 + e^t,\end{aligned}$$

can be written as

$$\vec{x}' = \begin{bmatrix} 2t & e^t \\ 1/t & -1 \end{bmatrix} \vec{x} + \begin{bmatrix} t^2 \\ e^t \end{bmatrix}.$$

We will mostly concentrate on equations that are not just linear, but are in fact *constant coefficient* equations. That is, the matrix P will be constant; it will not depend on t .

When $\vec{f} = \vec{0}$ (the zero vector), then we say the system is *homogeneous*. For homogeneous linear systems we have the principle of superposition, just like for single homogeneous equations.

Theorem 4.3.1 (Superposition)

Let $\vec{x}' = P\vec{x}$ be a linear homogeneous system of ODEs. Suppose that $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$ are n solutions of the equation and c_1, c_2, \dots, c_n are any constants, then

$$\vec{x} = c_1 \vec{x}_1 + c_2 \vec{x}_2 + \cdots + c_n \vec{x}_n, \quad (4.2)$$

is also a solution. Furthermore, if this is a system of n equations (P is $n \times n$), and $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$ are linearly independent, then every solution \vec{x} can be written as (4.2).

Linear independence for vector-valued functions is the same idea as for normal functions. The vector-valued functions $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$ are linearly independent when

$$c_1 \vec{x}_1 + c_2 \vec{x}_2 + \cdots + c_n \vec{x}_n = \vec{0}$$

has only the solution $c_1 = c_2 = \cdots = c_n = 0$, where the equation must hold for all t .

Example 4.3.1: $\vec{x}_1 = \begin{bmatrix} t^2 \\ t \end{bmatrix}$, $\vec{x}_2 = \begin{bmatrix} 0 \\ 1+t \end{bmatrix}$, $\vec{x}_3 = \begin{bmatrix} -t^2 \\ 1 \end{bmatrix}$ are linearly dependent because $\vec{x}_1 + \vec{x}_3 = \vec{x}_2$, and this holds for all t . So $c_1 = 1$, $c_2 = -1$, and $c_3 = 1$ above will work.

On the other hand if we change the example just slightly $\vec{x}_1 = \begin{bmatrix} t^2 \\ t \end{bmatrix}$, $\vec{x}_2 = \begin{bmatrix} 0 \\ t \end{bmatrix}$, $\vec{x}_3 = \begin{bmatrix} -t^2 \\ 1 \end{bmatrix}$, then the functions are linearly independent. First write $c_1 \vec{x}_1 + c_2 \vec{x}_2 + c_3 \vec{x}_3 = \vec{0}$ and note that it has to hold for all t . We get that

$$c_1 \vec{x}_1 + c_2 \vec{x}_2 + c_3 \vec{x}_3 = \begin{bmatrix} c_1 t^2 - c_3 t^2 \\ c_1 t + c_2 t + c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

In other words $c_1 t^2 - c_3 t^2 = 0$ and $c_1 t + c_2 t + c_3 = 0$. If we set $t = 0$, then the second equation becomes $c_3 = 0$. But then the first equation becomes $c_1 t^2 = 0$ for all t and so $c_1 = 0$. Thus

the second equation is just $c_2 t = 0$, which means $c_2 = 0$. So $c_1 = c_2 = c_3 = 0$ is the only solution and \vec{x}_1 , \vec{x}_2 , and \vec{x}_3 are linearly independent.

The linear combination $c_1 \vec{x}_1 + c_2 \vec{x}_2 + \cdots + c_n \vec{x}_n$ could always be written as

$$X(t) \vec{c},$$

where $X(t)$ is the matrix with columns $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$, and \vec{c} is the column vector with entries c_1, c_2, \dots, c_n . Assuming that $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$ are linearly independent and solutions to a given system of differential equations, the matrix-valued function $X(t)$ is called a *fundamental matrix*, or a *fundamental matrix solution*.

To solve nonhomogeneous first order linear systems, we use the same technique as we applied to solve single linear nonhomogeneous equations.

Theorem 4.3.2

Let $\vec{x}' = P\vec{x} + \vec{f}$ be a linear system of ODEs. Suppose \vec{x}_p is one particular solution. Then every solution can be written as

$$\vec{x} = \vec{x}_c + \vec{x}_p,$$

where \vec{x}_c is a solution to the associated homogeneous equation ($\vec{x}' = P\vec{x}$).

The procedure for systems is the same as for single equations. We find a particular solution to the nonhomogeneous equation, then we find the general solution to the associated homogeneous equation, and finally we add the two together.

Alright, suppose you have found the general solution of $\vec{x}' = P\vec{x} + \vec{f}$. Next suppose you are given an initial condition of the form

$$\vec{x}(t_0) = \vec{b}$$

for some fixed t_0 and a constant vector \vec{b} . Let $X(t)$ be a fundamental matrix solution of the associated homogeneous equation (i.e. columns of $X(t)$ are solutions). The general solution can be written as

$$\vec{x}(t) = X(t) \vec{c} + \vec{x}_p(t).$$

We are seeking a vector \vec{c} such that

$$\vec{b} = \vec{x}(t_0) = X(t_0) \vec{c} + \vec{x}_p(t_0).$$

In other words, we are solving for \vec{c} in the nonhomogeneous system of linear equations

$$X(t_0) \vec{c} = \vec{b} - \vec{x}_p(t_0).$$

Example 4.3.2: In § 4.1 we solved the system

$$\begin{aligned} x'_1 &= x_1, \\ x'_2 &= x_1 - x_2, \end{aligned}$$

with initial conditions $x_1(0) = 1$, $x_2(0) = 2$. Let us consider this problem in the language of this section.

The system is homogeneous, so $\vec{f}(t) = \vec{0}$. We write the system and the initial conditions as

$$\vec{x}' = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \vec{x}, \quad \vec{x}(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

We found the general solution is $x_1 = c_1 e^t$ and $x_2 = \frac{c_1}{2} e^t + c_2 e^{-t}$. Letting $c_1 = 1$ and $c_2 = 0$, we obtain the solution $\begin{bmatrix} e^t \\ (1/2)e^t \end{bmatrix}$. Letting $c_1 = 0$ and $c_2 = 1$, we obtain $\begin{bmatrix} 0 \\ e^{-t} \end{bmatrix}$. These two solutions are linearly independent, as can be seen by setting $t = 0$, and noting that the resulting constant vectors are linearly independent. In matrix notation, a fundamental matrix solution is, therefore,

$$X(t) = \begin{bmatrix} e^t & 0 \\ \frac{1}{2}e^t & e^{-t} \end{bmatrix}.$$

To solve the initial value problem we solve for \vec{c} in the equation

$$X(0) \vec{c} = \vec{b},$$

or in other words,

$$\begin{bmatrix} 1 & 0 \\ \frac{1}{2} & 1 \end{bmatrix} \vec{c} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

A single elementary row operation shows $\vec{c} = \begin{bmatrix} 1 \\ 3/2 \end{bmatrix}$. Our solution is

$$\vec{x}(t) = X(t) \vec{c} = \begin{bmatrix} e^t & 0 \\ \frac{1}{2}e^t & e^{-t} \end{bmatrix} \begin{bmatrix} 1 \\ \frac{3}{2} \end{bmatrix} = \begin{bmatrix} e^t & e^t \\ \frac{1}{2}e^t + \frac{3}{2}e^{-t} & e^{-t} \end{bmatrix}.$$

This new solution agrees with our previous solution from § 4.1.

4.3.1 Exercises

Exercise 4.3.1: Write the system $x'_1 = 2x_1 - 3tx_2 + \sin t$, $x'_2 = e^t x_1 + 3x_2 + \cos t$ in the form $\vec{x}' = P(t)\vec{x} + \vec{f}(t)$.

Exercise 4.3.2:* Write $x' = 3x - y + e^t$, $y' = tx$ in matrix notation.

Exercise 4.3.3:

a) Verify that the system $\vec{x}' = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix} \vec{x}$ has the two solutions $\begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{4t}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-2t}$.

b) Write down the general solution.

c) Write down the general solution in the form $x_1 = ?$, $x_2 = ?$ (i.e. write down a formula for each element of the solution).

Exercise 4.3.4: Verify that $\begin{bmatrix} 1 \\ 1 \end{bmatrix} e^t$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix} e^t$ are linearly independent. Hint: Just plug in $t = 0$.

Exercise 4.3.5:* Are $\begin{bmatrix} e^{2t} \\ e^t \end{bmatrix}$ and $\begin{bmatrix} e^t \\ e^{2t} \end{bmatrix}$ linearly independent? Justify.

Exercise 4.3.6: Verify that $\begin{bmatrix} 1 \\ 0 \end{bmatrix} e^t$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix} e^t$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{2t}$ are linearly independent. Hint: You must be a bit more tricky than in the previous exercise.

Exercise 4.3.7:* Are $\begin{bmatrix} \cosh(t) \\ 1 \end{bmatrix}$, $\begin{bmatrix} e^t \\ 1 \end{bmatrix}$, and $\begin{bmatrix} e^{-t} \\ 1 \end{bmatrix}$ linearly independent? Justify.

Exercise 4.3.8: Verify that $\begin{bmatrix} t \\ t^2 \end{bmatrix}$ and $\begin{bmatrix} t^3 \\ t^4 \end{bmatrix}$ are linearly independent.

Exercise 4.3.9: Take the system $x'_1 + x'_2 = x_1$, $x'_1 - x'_2 = x_2$.

- a) Write it in the form $A\vec{x}' = B\vec{x}$ for matrices A and B .
- b) Compute A^{-1} and use that to write the system in the form $\vec{x}' = P\vec{x}$.

Exercise 4.3.10:*

- a) Write $x'_1 = 2tx_2$, $x'_2 = 2tx_1$ in matrix notation.
- b) Solve and write the solution in matrix notation.

4.4 Eigenvalue method

Attribution: [JL], §3.4, 3.7.

Learning Objectives

After this section, you will be able to:

- Use the eigenvalue method to find straight-line solutions to constant-coefficient first order systems of ODE,
- Find general solutions to systems with real and distinct eigenvalues,
- Use Euler's formula to find a real-valued general solution to a first order system with complex eigenvalues,
- Find generalized eigenvectors to write a general solution to a first order system with repeated and defective eigenvalues, and
- Solve initial value problems from all of these cases once the general solution has been found.

In this section we will learn how to solve linear homogeneous constant coefficient systems of ODEs by the eigenvalue method. Suppose we have such a system

$$\vec{x}' = P\vec{x},$$

where P is a constant square matrix. We wish to adapt the method for the single constant coefficient equation by trying the function $e^{\lambda t}$. However, \vec{x} is a vector. So we try $\vec{x} = \vec{v}e^{\lambda t}$, where \vec{v} is an arbitrary constant vector. We plug this \vec{x} into the equation to get

$$\underbrace{\lambda \vec{v} e^{\lambda t}}_{\vec{x}'} = \underbrace{P\vec{v} e^{\lambda t}}_{P\vec{x}}.$$

We divide by $e^{\lambda t}$ and notice that we are looking for a scalar λ and a vector \vec{v} that satisfy the equation

$$\lambda \vec{v} = P\vec{v}.$$

This means that we are looking for an eigenvalue λ with corresponding eigenvector \vec{v} for the matrix P . When we can find these, we will get solutions to the original system of differential equations of the form

$$\vec{x}(t) = \vec{v}e^{\lambda t}.$$

We get the easiest route to solutions when the matrix P has all real eigenvalues and the eigenvalues are all distinct, and can extend to deal with the complications that arise from complex and repeated eigenvalues.

4.4.1 The eigenvalue method with distinct real eigenvalues

OK. We have the system of equations

$$\vec{x}' = P\vec{x}.$$

We find the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ of the matrix P , and corresponding eigenvectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$. Now we notice that the functions $\vec{v}_1 e^{\lambda_1 t}, \vec{v}_2 e^{\lambda_2 t}, \dots, \vec{v}_n e^{\lambda_n t}$ are solutions of the system of equations and hence $\vec{x} = c_1 \vec{v}_1 e^{\lambda_1 t} + c_2 \vec{v}_2 e^{\lambda_2 t} + \dots + c_n \vec{v}_n e^{\lambda_n t}$ is a solution.

Theorem 4.4.1

Take $\vec{x}' = P\vec{x}$. If P is an $n \times n$ constant matrix that has n distinct real eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$, then there exist n linearly independent corresponding eigenvectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$, and the general solution to $\vec{x}' = P\vec{x}$ can be written as

$$\vec{x} = c_1 \vec{v}_1 e^{\lambda_1 t} + c_2 \vec{v}_2 e^{\lambda_2 t} + \dots + c_n \vec{v}_n e^{\lambda_n t}.$$

The corresponding fundamental matrix solution is

$$X(t) = [\vec{v}_1 e^{\lambda_1 t} \quad \vec{v}_2 e^{\lambda_2 t} \quad \dots \quad \vec{v}_n e^{\lambda_n t}].$$

That is, $X(t)$ is the matrix whose j^{th} column is $\vec{v}_j e^{\lambda_j t}$.

Example 4.4.1: Consider the system

$$\vec{x}' = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \vec{x}.$$

Find the general solution.

Solution: Earlier, we found the eigenvalues are 1, 2, 3. We found the eigenvector $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ for the eigenvalue 3. Similarly we find the eigenvector $\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$ for the eigenvalue 1, and $\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$ for the eigenvalue 2 (exercise: check). Hence our general solution is

$$\vec{x} = c_1 \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} e^t + c_2 \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} e^{2t} + c_3 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} e^{3t} = \begin{bmatrix} c_1 e^t + c_3 e^{3t} \\ -c_1 e^t + c_2 e^{2t} + c_3 e^{3t} \\ -c_2 e^{2t} \end{bmatrix}.$$

In terms of a fundamental matrix solution,

$$\vec{x} = X(t) \vec{c} = \begin{bmatrix} e^t & 0 & e^{3t} \\ -e^t & e^{2t} & e^{3t} \\ 0 & -e^{2t} & 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}.$$

Exercise 4.4.1: Check that this \vec{x} really solves the system.

Overall, the process for finding the solution for real and distinct eigenvalues is to first find the eigenvalues and eigenvectors of the matrix P . Once we have these, we get n linearly independent solutions of the form $\vec{x}_i(t) = \vec{v}_i e^{\lambda_i t}$, so that the general solution is of the form

$$\vec{x}(t) = c_1 \vec{v}_1 e^{\lambda_1 t} + c_2 \vec{v}_2 e^{\lambda_2 t} + \dots + c_n \vec{v}_n e^{\lambda_n t}.$$

Then, if we need to solve for an initial condition, we figure out the coefficients c_1, c_2, \dots, c_n to satisfy this condition.

Note: If we write a single homogeneous linear constant coefficient n^{th} order equation as a first order system (as we did in § 4.1), then the eigenvalue equation

$$\det(P - \lambda I) = 0$$

is essentially the same as the characteristic equation we got in § 2.1 and § 2.7.

4.4.2 Complex eigenvalues

A matrix may very well have complex eigenvalues even if all the entries are real. Take, for example,

$$\vec{x}' = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \vec{x}.$$

Let us compute the eigenvalues of the matrix $P = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$.

$$\det(P - \lambda I) = \det \left(\begin{bmatrix} 1 - \lambda & 1 \\ -1 & 1 - \lambda \end{bmatrix} \right) = (1 - \lambda)^2 + 1 = \lambda^2 - 2\lambda + 2 = 0.$$

Thus $\lambda = 1 \pm i$. Corresponding eigenvectors are also complex. Start with $\lambda = 1 - i$.

$$(P - (1 - i)I)\vec{v} = \vec{0},$$

$$\begin{bmatrix} i & 1 \\ -1 & i \end{bmatrix} \vec{v} = \vec{0}.$$

The equations $iv_1 + v_2 = 0$ and $-v_1 + iv_2 = 0$ are multiples of each other. So we only need to consider one of them. After picking $v_2 = 1$, for example, we have an eigenvector $\vec{v} = \begin{bmatrix} i \\ 1 \end{bmatrix}$. In similar fashion we find that $\begin{bmatrix} -i \\ 1 \end{bmatrix}$ is an eigenvector corresponding to the eigenvalue $1 + i$.

We could write the solution as

$$\vec{x} = c_1 \begin{bmatrix} i \\ 1 \end{bmatrix} e^{(1-i)t} + c_2 \begin{bmatrix} -i \\ 1 \end{bmatrix} e^{(1+i)t} = \begin{bmatrix} c_1 ie^{(1-i)t} - c_2 ie^{(1+i)t} \\ c_1 e^{(1-i)t} + c_2 e^{(1+i)t} \end{bmatrix}.$$

We would then need to look for complex values c_1 and c_2 to solve any initial conditions. It is perhaps not completely clear that we get a real solution. After solving for c_1 and c_2 , we could use [Euler's formula](#) and do the whole song and dance we did before, but we will not. We will apply the formula in a smarter way first to find independent real solutions.

We claim that we did not have to look for a second eigenvector (nor for the second eigenvalue). All complex eigenvalues come in pairs (because the matrix P is real).

First a small detour. The real part of a complex number z can be computed as $\frac{z + \bar{z}}{2}$, where the bar above z means $\overline{a + ib} = a - ib$. This operation is called the *complex conjugate*. If a is a real number, then $\bar{a} = a$. Similarly we bar whole vectors or matrices by taking the complex conjugate of every entry. Suppose a matrix P is real. Then $\bar{P} = P$, and so $\bar{P}\vec{x} = \bar{P}\bar{\vec{x}} = P\vec{x}$. Also the complex conjugate of 0 is still 0, therefore,

$$\vec{0} = \bar{\vec{0}} = \overline{(P - \lambda I)\vec{v}} = (P - \bar{\lambda}I)\bar{\vec{v}}.$$

In other words, if $\lambda = a + ib$ is an eigenvalue, then so is $\bar{\lambda} = a - ib$. And if \vec{v} is an eigenvector corresponding to the eigenvalue λ , then $\bar{\vec{v}}$ is an eigenvector corresponding to the eigenvalue $\bar{\lambda}$.

Suppose $a + ib$ is a complex eigenvalue of P , and \vec{v} is a corresponding eigenvector. Then

$$\vec{x}_1 = \vec{v}e^{(a+ib)t}$$

is a solution (complex-valued) of $\vec{x}' = P\vec{x}$. **Euler's formula** shows that $\overline{e^{a+ib}} = e^{a-ib}$, and so

$$\vec{x}_2 = \bar{\vec{x}}_1 = \bar{\vec{v}}e^{(a-ib)t}$$

is also a solution. As \vec{x}_1 and \vec{x}_2 are solutions, the function

$$\vec{x}_3 = \operatorname{Re} \vec{x}_1 = \operatorname{Re} \vec{v}e^{(a+ib)t} = \frac{\vec{x}_1 + \bar{\vec{x}}_1}{2} = \frac{\vec{x}_1 + \vec{x}_2}{2} = \frac{1}{2}\vec{x}_1 + \frac{1}{2}\vec{x}_2$$

is also a solution. And \vec{x}_3 is real-valued! Similarly as $\operatorname{Im} z = \frac{z-\bar{z}}{2i}$ is the imaginary part, we find that

$$\vec{x}_4 = \operatorname{Im} \vec{x}_1 = \frac{\vec{x}_1 - \bar{\vec{x}}_1}{2i} = \frac{\vec{x}_1 - \vec{x}_2}{2i}.$$

is also a real-valued solution. It turns out that \vec{x}_3 and \vec{x}_4 are linearly independent. We will use **Euler's formula** to separate out the real and imaginary part.

Returning to our problem,

$$\vec{x}_1 = \begin{bmatrix} i \\ 1 \end{bmatrix} e^{(1-i)t} = \begin{bmatrix} i \\ 1 \end{bmatrix} (e^t \cos t - ie^t \sin t) = \begin{bmatrix} ie^t \cos t + e^t \sin t \\ e^t \cos t - ie^t \sin t \end{bmatrix} = \begin{bmatrix} e^t \sin t \\ e^t \cos t \end{bmatrix} + i \begin{bmatrix} e^t \cos t \\ -e^t \sin t \end{bmatrix}.$$

Then

$$\operatorname{Re} \vec{x}_1 = \begin{bmatrix} e^t \sin t \\ e^t \cos t \end{bmatrix}, \quad \text{and} \quad \operatorname{Im} \vec{x}_1 = \begin{bmatrix} e^t \cos t \\ -e^t \sin t \end{bmatrix},$$

are the two real-valued linearly independent solutions we seek.

Exercise 4.4.2: Check that these really are solutions.

The general solution is

$$\vec{x} = c_1 \begin{bmatrix} e^t \sin t \\ e^t \cos t \end{bmatrix} + c_2 \begin{bmatrix} e^t \cos t \\ -e^t \sin t \end{bmatrix} = \begin{bmatrix} c_1 e^t \sin t + c_2 e^t \cos t \\ c_1 e^t \cos t - c_2 e^t \sin t \end{bmatrix}.$$

This solution is real-valued for real c_1 and c_2 . We now solve for any initial conditions we may have.

Let us summarize as a theorem.

Theorem 4.4.2

Let P be a real-valued constant matrix. If P has a complex eigenvalue $a + ib$ and a corresponding eigenvector \vec{v} , then P also has a complex eigenvalue $a - ib$ with a corresponding eigenvector $\bar{\vec{v}}$. Furthermore, $\vec{x}' = P\vec{x}$ has two linearly independent real-valued solutions

$$\vec{x}_1 = \operatorname{Re} \vec{v}e^{(a+ib)t}, \quad \text{and} \quad \vec{x}_2 = \operatorname{Im} \vec{v}e^{(a+ib)t}.$$

The main point here is that the real and imaginary parts of these complex solutions are the real-valued independent solutions that we seek. Compare this to Theorem 2.2.2 in § 2.2, where we saw that the same idea worked for second order equation with complex roots.

For each pair of complex eigenvalues $a + ib$ and $a - ib$, we get two real-valued linearly independent solutions. We then go on to the next eigenvalue, which is either a real eigenvalue or another complex eigenvalue pair. If we have n distinct eigenvalues (real or complex), then we end up with n linearly independent solutions. If we had only two equations ($n = 2$) as in the example above, then once we found two solutions we are finished, and our general solution is

$$\vec{x} = c_1 \vec{x}_1 + c_2 \vec{x}_2 = c_1 (\operatorname{Re} \vec{v} e^{(a+ib)t}) + c_2 (\operatorname{Im} \vec{v} e^{(a+ib)t}).$$

We can now find a real-valued general solution to any homogeneous system where the matrix has distinct eigenvalues. When we have repeated eigenvalues, matters get a bit more complicated.

4.4.3 Repeated Eigenvalues

It may happen that a matrix A has some “repeated” eigenvalues. That is, the characteristic equation $\det(A - \lambda I) = 0$ may have repeated roots. This is actually unlikely to happen for a random matrix. If we take a small perturbation of A (we change the entries of A slightly), we get a matrix with distinct eigenvalues. As any system we want to solve in practice is an approximation to reality anyway, it is not absolutely indispensable to know how to solve these corner cases. On the other hand, these cases do come up in applications from time to time. Furthermore, if we have distinct but very close eigenvalues, the behavior is similar to that of repeated eigenvalues, and so understanding that case will give us insight into what is going on.

Geometric multiplicity

Take the diagonal matrix

$$A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}.$$

A has an eigenvalue 3 of multiplicity 2. We call the multiplicity of the eigenvalue in the characteristic equation the *algebraic multiplicity*. In this case, there also exist 2 linearly independent eigenvectors, $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ corresponding to the eigenvalue 3. This means that the so-called *geometric multiplicity* of this eigenvalue is also 2.

In all the theorems where we required a matrix to have n distinct eigenvalues, we only really needed to have n linearly independent eigenvectors. For example, $\vec{x}' = A\vec{x}$ has the general solution

$$\vec{x} = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{3t} + c_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{3t}.$$

Let us restate the theorem about real eigenvalues. In the following theorem we will repeat eigenvalues according to (algebraic) multiplicity. So for the matrix A above, we would say that it has eigenvalues 3 and 3.

Theorem 4.4.3

Suppose the $n \times n$ matrix P has n real eigenvalues (not necessarily distinct), $\lambda_1, \lambda_2, \dots, \lambda_n$, and there are n linearly independent corresponding eigenvectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$. Then the general solution to $\vec{x}' = P\vec{x}$ can be written as

$$\vec{x} = c_1 \vec{v}_1 e^{\lambda_1 t} + c_2 \vec{v}_2 e^{\lambda_2 t} + \cdots + c_n \vec{v}_n e^{\lambda_n t}.$$

The *geometric multiplicity* of an eigenvalue of algebraic multiplicity n is equal to the number of corresponding linearly independent eigenvectors. The geometric multiplicity is always less than or equal to the algebraic multiplicity. The theorem handles the case when these two multiplicities are equal for all eigenvalues. If for an eigenvalue the geometric multiplicity is equal to the algebraic multiplicity, then we say the eigenvalue is *complete*.

In other words, the hypothesis of the theorem could be stated as saying that if all the eigenvalues of P are complete, then there are n linearly independent eigenvectors and thus we have the given general solution.

If the geometric multiplicity of an eigenvalue is 2 or greater, then the set of linearly independent eigenvectors is not unique up to multiples as it was before. For example, for the diagonal matrix $A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$ we could also pick eigenvectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$, or in fact any pair of two linearly independent vectors. The number of linearly independent eigenvectors corresponding to λ is the number of free variables we obtain when solving $A\vec{v} = \lambda\vec{v}$. We pick specific values for those free variables to obtain eigenvectors. If you pick different values, you may get different eigenvectors.

Defective eigenvalues

If an $n \times n$ matrix has less than n linearly independent eigenvectors, it is said to be *deficient*. Then there is at least one eigenvalue with an algebraic multiplicity that is higher than its geometric multiplicity. We call this eigenvalue *defective* and the difference between the two multiplicities we call the *defect*.

Example 4.4.2: The matrix

$$\begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$$

has an eigenvalue 3 of algebraic multiplicity 2. Let us try to compute eigenvectors.

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \vec{0}.$$

We must have that $v_2 = 0$. Hence any eigenvector is of the form $\begin{bmatrix} v_1 \\ 0 \end{bmatrix}$. Any two such vectors are linearly dependent, and hence the geometric multiplicity of the eigenvalue is 1. Therefore, the defect is 1, and we can no longer apply the eigenvalue method directly to a system of ODEs with such a coefficient matrix.

Roughly, the key observation is that if λ is an eigenvalue of A of algebraic multiplicity m , then we can find certain m linearly independent vectors solving $(A - \lambda I)^k \vec{v} = \vec{0}$ for various powers k . We will call these *generalized eigenvectors*.

Let us continue with the example $A = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}$ and the equation $\vec{x}' = A\vec{x}$. We found an eigenvalue $\lambda = 3$ of (algebraic) multiplicity 2 and defect 1. We found one eigenvector $\vec{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. We have one solution

$$\vec{x}_1 = \vec{v}e^{3t} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{3t}.$$

We are now stuck, we get no other solutions from standard eigenvectors. But we need two linearly independent solutions to find the general solution of the equation.

Let us try (in the spirit of repeated roots of the characteristic equation for a single equation) another solution of the form

$$\vec{x}_2 = (\vec{v}_2 + \vec{v}_1 t) e^{3t}.$$

We differentiate to get

$$\vec{x}'_2 = \vec{v}_1 e^{3t} + 3(\vec{v}_2 + \vec{v}_1 t) e^{3t} = (3\vec{v}_2 + \vec{v}_1) e^{3t} + 3\vec{v}_1 t e^{3t}.$$

As we are assuming that \vec{x}_2 is a solution, \vec{x}'_2 must equal $A\vec{x}_2$. So let's compute $A\vec{x}_2$:

$$A\vec{x}_2 = A(\vec{v}_2 + \vec{v}_1 t) e^{3t} = A\vec{v}_2 e^{3t} + A\vec{v}_1 t e^{3t}.$$

By looking at the coefficients of e^{3t} and $t e^{3t}$ we see $3\vec{v}_2 + \vec{v}_1 = A\vec{v}_2$ and $3\vec{v}_1 = A\vec{v}_1$. This means that

$$(A - 3I)\vec{v}_2 = \vec{v}_1, \quad \text{and} \quad (A - 3I)\vec{v}_1 = \vec{0}.$$

Therefore, \vec{x}_2 is a solution if these two equations are satisfied. The second equation is satisfied if \vec{v}_1 is an eigenvector, and we found the eigenvector above, so let $\vec{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. So, if we can find a \vec{v}_2 that solves $(A - 3I)\vec{v}_2 = \vec{v}_1$, then we are done. This is just a bunch of linear equations to solve and we are by now very good at that. Let us solve $(A - 3I)\vec{v}_2 = \vec{v}_1$. Write

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

By inspection we see that letting $a = 0$ (a could be anything in fact) and $b = 1$ does the job. Hence we can take $\vec{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Our general solution to $\vec{x}' = A\vec{x}$ is

$$\vec{x} = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{3t} + c_2 \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} t \right) e^{3t} = \begin{bmatrix} c_1 e^{3t} + c_2 t e^{3t} \\ c_2 e^{3t} \end{bmatrix}.$$

Let us check that we really do have the solution. First $x'_1 = c_1 3e^{3t} + c_2 e^{3t} + 3c_2 t e^{3t} = 3x_1 + x_2$. Good. Now $x'_2 = 3c_2 t e^{3t} = 3x_2$. Good.

In the example, if we plug $(A - 3I)\vec{v}_2 = \vec{v}_1$ into $(A - 3I)\vec{v}_1 = \vec{0}$ we find

$$(A - 3I)(A - 3I)\vec{v}_2 = \vec{0}, \quad \text{or} \quad (A - 3I)^2 \vec{v}_2 = \vec{0}.$$

Furthermore, if $(A - 3I)\vec{w} \neq \vec{0}$, then $(A - 3I)\vec{w}$ is an eigenvector, a multiple of \vec{v}_1 . In this 2×2 case $(A - 3I)^2$ is just the zero matrix (exercise). So any vector \vec{w} solves $(A - 3I)^2 \vec{w} = \vec{0}$ and we just need a \vec{w} such that $(A - 3I)\vec{w} \neq \vec{0}$. Then we could use \vec{w} for \vec{v}_2 , and $(A - 3I)\vec{w}$ for \vec{v}_1 .

Note that the system $\vec{x}' = A\vec{x}$ has a simpler solution since A is a so-called *upper triangular matrix*, that is every entry below the diagonal is zero. In particular, the equation for x_2 does not depend on x_1 . Mind you, not every defective matrix is triangular.

Exercise 4.4.3: Solve $\vec{x}' = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix} \vec{x}$ by first solving for x_2 and then for x_1 independently. Check that you got the same solution as we did above.

Let us describe the general algorithm. Suppose that λ is an eigenvalue of multiplicity 2, defect 1. First find an eigenvector \vec{v}_1 of λ . That is, \vec{v}_1 solves $(A - \lambda I)\vec{v}_1 = \vec{0}$. Then, find a vector \vec{v}_2 such that

$$(A - \lambda I)\vec{v}_2 = \vec{v}_1.$$

This gives us two linearly independent solutions

$$\begin{aligned}\vec{x}_1 &= \vec{v}_1 e^{\lambda t}, \\ \vec{x}_2 &= (\vec{v}_2 + \vec{v}_1 t) e^{\lambda t}.\end{aligned}$$

Example 4.4.3: Consider the system

$$\vec{x}' = \begin{bmatrix} 2 & -5 & 0 \\ 0 & 2 & 0 \\ -1 & 4 & 1 \end{bmatrix} \vec{x}.$$

Find the general solution to this system using eigenvalues and eigenvectors.

Solution: Compute the eigenvalues,

$$0 = \det(A - \lambda I) = \det \left(\begin{bmatrix} 2 - \lambda & -5 & 0 \\ 0 & 2 - \lambda & 0 \\ -1 & 4 & 1 - \lambda \end{bmatrix} \right) = (2 - \lambda)^2(1 - \lambda).$$

The eigenvalues are 1 and 2, where 2 has multiplicity 2. We leave it to the reader to find that $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ is an eigenvector for the eigenvalue $\lambda = 1$.

Let's focus on $\lambda = 2$. We compute eigenvectors:

$$\vec{0} = (A - 2I)\vec{v} = \begin{bmatrix} 0 & -5 & 0 \\ 0 & 0 & 0 \\ -1 & 4 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

The first equation says that $v_2 = 0$, so the last equation is $-v_1 - v_3 = 0$. Let v_3 be the free variable to find that $v_1 = -v_3$. Perhaps let $v_3 = -1$ to find an eigenvector $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$. Problem is that setting v_3 to anything else just gets multiples of this vector and so we have a defect of 1. Let \vec{v}_1 be the eigenvector and let's look for a generalized eigenvector \vec{v}_2 :

$$(A - 2I)\vec{v}_2 = \vec{v}_1,$$

or

$$\begin{bmatrix} 0 & -5 & 0 \\ 0 & 0 & 0 \\ -1 & 4 & -1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix},$$

where we used a, b, c as components of \vec{v}_2 for simplicity. The first equation says $-5b = 1$ so $b = -1/5$. The second equation says nothing. The last equation is $-a + 4b - c = -1$, or

$a + \frac{4}{5} + c = 1$, or $a + c = \frac{1}{5}$. We let c be the free variable and we choose $c = 0$. We find $\vec{v}_2 = \begin{bmatrix} 1/5 \\ -1/5 \\ 0 \end{bmatrix}$.

The general solution is therefore,

$$\vec{x} = c_1 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} e^t + c_2 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} e^{2t} + c_3 \left(\begin{bmatrix} 1/5 \\ -1/5 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} t \right) e^{2t}.$$

]

This machinery can also be generalized to higher multiplicities and higher defects. We will not go over this method in detail, but let us just sketch the ideas. Suppose that A has an eigenvalue λ of multiplicity m . We find vectors such that

$$(A - \lambda I)^k \vec{v} = \vec{0}, \quad \text{but} \quad (A - \lambda I)^{k-1} \vec{v} \neq \vec{0}.$$

Such vectors are called *generalized eigenvectors* (then $\vec{v}_1 = (A - \lambda I)^{k-1} \vec{v}$ is an eigenvector). For the eigenvector \vec{v}_1 there is a chain of generalized eigenvectors \vec{v}_2 through \vec{v}_k such that:

$$\begin{aligned} (A - \lambda I) \vec{v}_1 &= \vec{0}, \\ (A - \lambda I) \vec{v}_2 &= \vec{v}_1, \\ &\vdots \\ (A - \lambda I) \vec{v}_k &= \vec{v}_{k-1}. \end{aligned}$$

Really once you find the \vec{v}_k such that $(A - \lambda I)^k \vec{v}_k = \vec{0}$ but $(A - \lambda I)^{k-1} \vec{v}_k \neq \vec{0}$, you find the entire chain since you can compute the rest, $\vec{v}_{k-1} = (A - \lambda I) \vec{v}_k$, $\vec{v}_{k-2} = (A - \lambda I) \vec{v}_{k-1}$, etc. We form the linearly independent solutions

$$\begin{aligned} \vec{x}_1 &= \vec{v}_1 e^{\lambda t}, \\ \vec{x}_2 &= (\vec{v}_2 + \vec{v}_1 t) e^{\lambda t}, \\ &\vdots \\ \vec{x}_k &= \left(\vec{v}_k + \vec{v}_{k-1} t + \vec{v}_{k-2} \frac{t^2}{2} + \cdots + \vec{v}_2 \frac{t^{k-2}}{(k-2)!} + \vec{v}_1 \frac{t^{k-1}}{(k-1)!} \right) e^{\lambda t}. \end{aligned}$$

Recall that $k! = 1 \cdot 2 \cdot 3 \cdots (k-1) \cdot k$ is the factorial. If you have an eigenvalue of geometric multiplicity ℓ , you will have to find ℓ such chains (some of them might be short: just the single eigenvector equation). We go until we form m linearly independent solutions where m is the algebraic multiplicity. We don't quite know which specific eigenvectors go with which chain, so start by finding \vec{v}_k first for the longest possible chain and go from there.

For example, if λ is an eigenvalue of A of algebraic multiplicity 3 and defect 2, then solve

$$(A - \lambda I) \vec{v}_1 = \vec{0}, \quad (A - \lambda I) \vec{v}_2 = \vec{v}_1, \quad (A - \lambda I) \vec{v}_3 = \vec{v}_2.$$

That is, find \vec{v}_3 such that $(A - \lambda I)^3 \vec{v}_3 = \vec{0}$, but $(A - \lambda I)^2 \vec{v}_3 \neq \vec{0}$. Then you are done as $\vec{v}_2 = (A - \lambda I) \vec{v}_3$ and $\vec{v}_1 = (A - \lambda I) \vec{v}_2$. The 3 linearly independent solutions are

$$\vec{x}_1 = \vec{v}_1 e^{\lambda t}, \quad \vec{x}_2 = (\vec{v}_2 + \vec{v}_1 t) e^{\lambda t}, \quad \vec{x}_3 = \left(\vec{v}_3 + \vec{v}_2 t + \vec{v}_1 \frac{t^2}{2} \right) e^{\lambda t}.$$

If on the other hand A has an eigenvalue λ of algebraic multiplicity 3 and defect 1, then solve

$$(A - \lambda I)\vec{v}_1 = \vec{0}, \quad (A - \lambda I)\vec{v}_2 = \vec{0}, \quad (A - \lambda I)\vec{v}_3 = \vec{v}_2.$$

Here \vec{v}_1 and \vec{v}_2 are actual honest eigenvectors, and \vec{v}_3 is a generalized eigenvector. So there are two chains. To solve, first find a \vec{v}_3 such that $(A - \lambda I)^2\vec{v}_3 = \vec{0}$, but $(A - \lambda I)\vec{v}_3 \neq \vec{0}$. Then $\vec{v}_2 = (A - \lambda I)\vec{v}_3$ is going to be an eigenvector. Then solve for an eigenvector \vec{v}_1 that is linearly independent from \vec{v}_2 . You get 3 linearly independent solutions

$$\vec{x}_1 = \vec{v}_1 e^{\lambda t}, \quad \vec{x}_2 = \vec{v}_2 e^{\lambda t}, \quad \vec{x}_3 = (\vec{v}_3 + \vec{v}_2 t) e^{\lambda t}.$$

4.4.4 Exercises

Exercise 4.4.4:

- a) Find the general solution of $x'_1 = 2x_1, x'_2 = 3x_2$ using the eigenvalue method (first write the system in the form $\vec{x}' = A\vec{x}$).
- b) Solve the system by solving each equation separately and verify you get the same general solution.

Exercise 4.4.5: Find the general solution of $x'_1 = 3x_1 + x_2, x'_2 = 2x_1 + 4x_2$ using the eigenvalue method.

Exercise 4.4.6:* Solve $x'_1 = x_2, x'_2 = x_1$ using the eigenvalue method.

Exercise 4.4.7: Find the general solution of $x'_1 = x_1 - 2x_2, x'_2 = 2x_1 + x_2$ using the eigenvalue method. Do not use complex exponentials in your solution.

Exercise 4.4.8:* Solve $x'_1 = x_2, x'_2 = -x_1$ using the eigenvalue method.

Exercise 4.4.9:

- a) Compute eigenvalues and eigenvectors of $A = \begin{bmatrix} 9 & -2 & -6 \\ -8 & 3 & 6 \\ 10 & -2 & -6 \end{bmatrix}$.

b) Find the general solution of $\vec{x}' = A\vec{x}$.

Exercise 4.4.10:*

- a) Compute eigenvalues and eigenvectors of $A = \begin{bmatrix} 1 & 0 & 3 \\ -1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix}$.

b) Solve the system $\vec{x}' = A\vec{x}$.

Exercise 4.4.11: Compute eigenvalues and eigenvectors of $\begin{bmatrix} -2 & -1 & -1 \\ 3 & 2 & 1 \\ -3 & -1 & 0 \end{bmatrix}$.

Exercise 4.4.12: Let a, b, c, d, e, f be numbers. Find the eigenvalues of $\begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix}$.

Exercise 4.4.13: Let $A = \begin{bmatrix} 5 & -3 \\ 3 & -1 \end{bmatrix}$. Find the general solution of $\vec{x}' = A\vec{x}$.

Exercise 4.4.14:*

- a) Compute eigenvalues and eigenvectors of $A = \begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}$.
 b) Solve the system $\vec{x}' = A\vec{x}$.

Exercise 4.4.15: Let $A = \begin{bmatrix} 5 & -4 & 4 \\ 0 & 3 & 0 \\ -2 & 4 & -1 \end{bmatrix}$.

- a) What are the eigenvalues?
 b) What is/are the defect(s) of the eigenvalue(s)?
 c) Find the general solution of $\vec{x}' = A\vec{x}$.

Exercise 4.4.16:* Let $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$.

- a) What are the eigenvalues?
 b) What is/are the defect(s) of the eigenvalue(s)?
 c) Find the general solution of $\vec{x}' = A\vec{x}$.

Exercise 4.4.17: Let $A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$.

- a) What are the eigenvalues?
 b) What is/are the defect(s) of the eigenvalue(s)?
 c) Find the general solution of $\vec{x}' = A\vec{x}$ in two different ways and verify you get the same answer.

Exercise 4.4.18:* Let $A = \begin{bmatrix} 1 & 3 & 3 \\ 1 & 1 & 0 \\ -1 & 1 & 2 \end{bmatrix}$.

- a) What are the eigenvalues?
 b) What is/are the defect(s) of the eigenvalue(s)?
 c) Find the general solution of $\vec{x}' = A\vec{x}$.

Exercise 4.4.19: Let $A = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 2 & 2 \\ -4 & 4 & 7 \end{bmatrix}$.

- a) What are the eigenvalues?
 b) What is/are the defect(s) of the eigenvalue(s)?
 c) Find the general solution of $\vec{x}' = A\vec{x}$.

Exercise 4.4.20:* Let $A = \begin{bmatrix} 2 & 0 & 0 \\ -1 & -1 & 9 \\ 0 & -1 & 5 \end{bmatrix}$.

- a) What are the eigenvalues?
 b) What is/are the defect(s) of the eigenvalue(s)?
 c) Find the general solution of $\vec{x}' = A\vec{x}$.

Exercise 4.4.21: Let $A = \begin{bmatrix} 0 & 4 & -2 \\ -1 & -4 & 1 \\ 0 & 0 & -2 \end{bmatrix}$.

- a) What are the eigenvalues?
- b) What is/are the defect(s) of the eigenvalue(s)?
- c) Find the general solution of $\vec{x}' = A\vec{x}$.

Exercise 4.4.22: Let $A = \begin{bmatrix} 2 & 1 & -1 \\ -1 & 0 & 2 \\ -1 & -2 & 4 \end{bmatrix}$.

- a) What are the eigenvalues?
- b) What is/are the defect(s) of the eigenvalue(s)?
- c) Find the general solution of $\vec{x}' = A\vec{x}$.

Exercise 4.4.23: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} -3 & 0 \\ 3 & -4 \end{bmatrix} \vec{x} \quad \vec{x}(0) = \begin{bmatrix} -1 \\ 2 \end{bmatrix}.$$

Exercise 4.4.24: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} 1 & -3 \\ 2 & 6 \end{bmatrix} \vec{x} \quad \vec{x}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Exercise 4.4.25: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} 3 & -1 \\ 4 & 3 \end{bmatrix} \vec{x} \quad \vec{x}(0) = \begin{bmatrix} 2 \\ -1 \end{bmatrix}.$$

Exercise 4.4.26: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} -3 & 2 \\ 0 & -3 \end{bmatrix} \vec{x} \quad \vec{x}(0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Exercise 4.4.27: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} 7 & 4 & 0 \\ -8 & -5 & 0 \\ 17 & 7 & -2 \end{bmatrix} \vec{x} \quad \vec{x}(0) = \begin{bmatrix} -3 \\ 2 \\ 2 \end{bmatrix}.$$

Exercise 4.4.28: Suppose that A is a 2×2 matrix with a repeated eigenvalue λ . Suppose that there are two linearly independent eigenvectors. Show that $A = \lambda I$.

Exercise 4.4.29:* Let $A = \begin{bmatrix} a & a \\ b & c \end{bmatrix}$, where a , b , and c are unknowns. Suppose that 5 is a doubled eigenvalue of defect 1, and suppose that $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is a corresponding eigenvector. Find A and show that there is only one such matrix A .

4.5 Two-dimensional systems and their vector fields

Attribution: [JL], §3.5.

Learning Objectives

After this section, you will be able to:

- Visualize and sketch the behavior of a two dimensional system based on the eigenvalues and eigenvectors.

Let us take a moment to talk about constant coefficient linear homogeneous systems in the plane. Much intuition can be obtained by studying this simple case. Suppose we use coordinates (x, y) for the plane as usual, and suppose $P = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is a 2×2 matrix. Consider the system

$$\begin{bmatrix} x \\ y \end{bmatrix}' = P \begin{bmatrix} x \\ y \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} x \\ y \end{bmatrix}' = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}. \quad (4.3)$$

The system is autonomous (compare this section to § 1.7) and so we can draw a vector field (see the end of § 4.1). We will be able to visually tell what the vector field looks like and how the solutions behave, once we find the eigenvalues and eigenvectors of the matrix P . For this section, we assume that P has two eigenvalues and two corresponding eigenvectors.

Case 1. Suppose that the eigenvalues of P are real and positive. We find two corresponding eigenvectors and plot them in the plane. For example, take the matrix $\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$. The eigenvalues are 1 and 2 and corresponding eigenvectors are $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. See Figure 4.4.

Suppose the point (x, y) is on the line determined by an eigenvector \vec{v} for an eigenvalue λ . That is, $\begin{bmatrix} x \\ y \end{bmatrix} = \alpha \vec{v}$ for some scalar α . Then

$$\begin{bmatrix} x \\ y \end{bmatrix}' = P \begin{bmatrix} x \\ y \end{bmatrix} = P(\alpha \vec{v}) = \alpha(P\vec{v}) = \alpha \lambda \vec{v}.$$

The derivative is a multiple of \vec{v} and hence points along the line determined by \vec{v} . As $\lambda > 0$, the derivative points in the direction of \vec{v} when α is positive and in the opposite direction when α is negative. Let us draw the lines determined by the eigenvectors, and let us draw arrows on the lines to indicate the directions. See Figure 4.5 on the next page.

We fill in the rest of the arrows for the vector field and we also draw a few solutions. See Figure 4.6 on the facing page. The picture looks like a source with arrows coming out from the origin. Hence we call this type of picture a *source* or sometimes an *unstable node*.

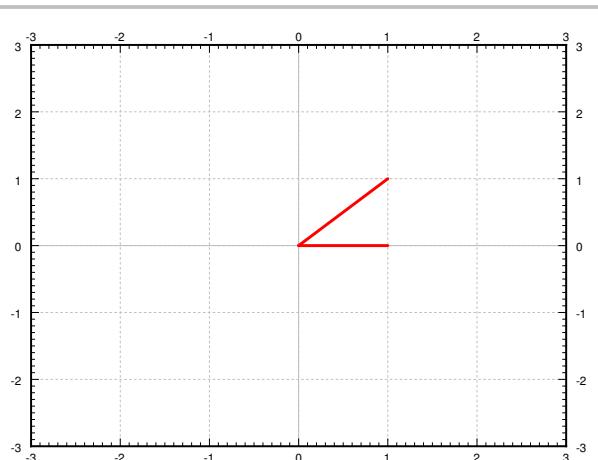


Figure 4.4: Eigenvectors of P .

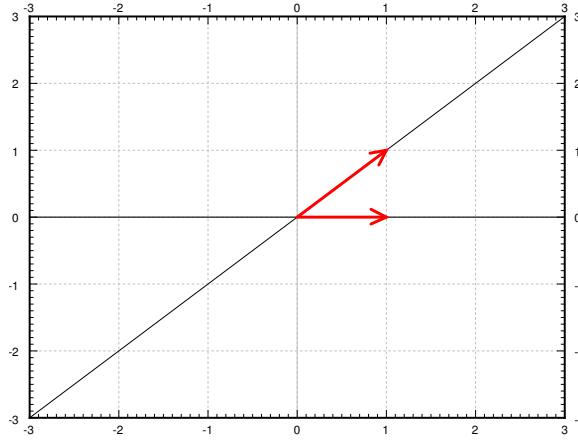
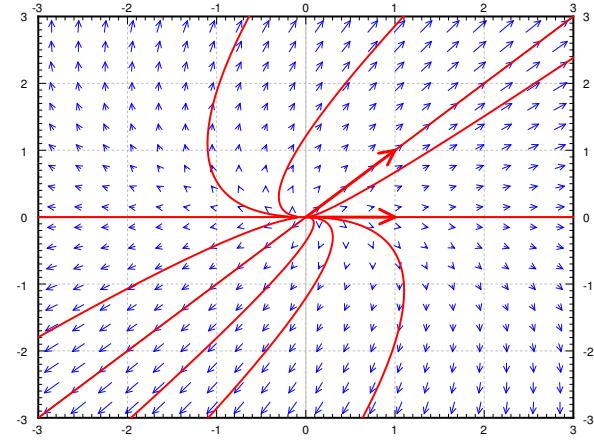
Figure 4.5: Eigenvectors of P with directions.

Figure 4.6: Example source vector field with eigenvectors and solutions.

Case 2. Suppose both eigenvalues are negative. For example, take the negation of the matrix in case 1, $\begin{bmatrix} -1 & -1 \\ 0 & -2 \end{bmatrix}$. The eigenvalues are -1 and -2 and corresponding eigenvectors are the same, $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. The calculation and the picture are almost the same. The only difference is that the eigenvalues are negative and hence all arrows are reversed. We get the picture in Figure 4.7. We call this kind of picture a *sink* or a *asymptotically stable node*.

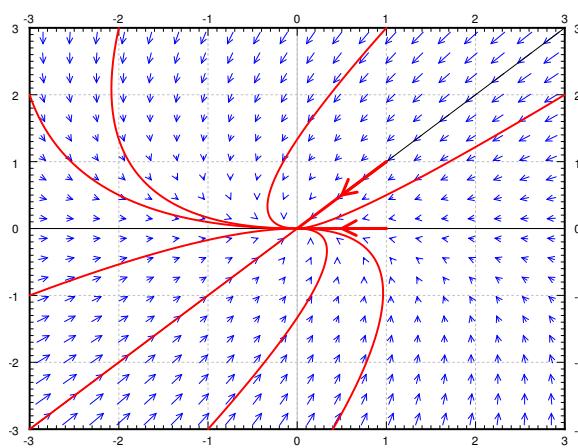


Figure 4.7: Example sink vector field with eigenvectors and solutions.

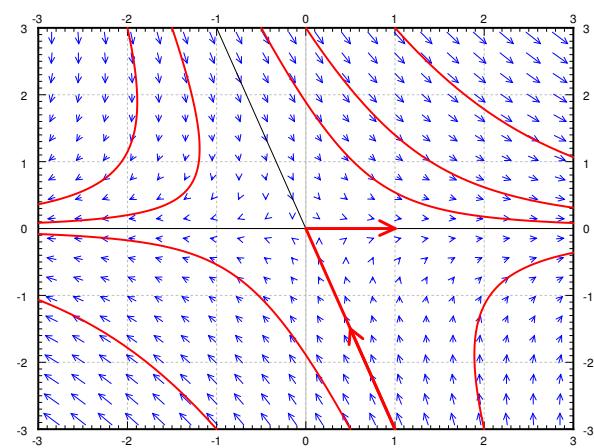


Figure 4.8: Example saddle vector field with eigenvectors and solutions.

Case 3. Suppose one eigenvalue is positive and one is negative. For example the matrix $\begin{bmatrix} 1 & 1 \\ 0 & -2 \end{bmatrix}$. The eigenvalues are 1 and -2 and corresponding eigenvectors are $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -3 \end{bmatrix}$. We reverse the arrows on one line (corresponding to the negative eigenvalue) and we obtain the picture in Figure 4.8. We call this picture a *saddle point*.

For the next three cases we will assume the eigenvalues are complex. In this case the eigenvectors are also complex and we cannot just plot them in the plane.

Case 4. Suppose the eigenvalues are purely imaginary. That is, suppose the eigenvalues are $\pm ib$. For example, let $P = \begin{bmatrix} 0 & 1 \\ -4 & 0 \end{bmatrix}$. The eigenvalues turn out to be $\pm 2i$ and eigenvectors are $\begin{bmatrix} 1 \\ 2i \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -2i \end{bmatrix}$. Consider the eigenvalue $2i$ and its eigenvector $\begin{bmatrix} 1 \\ 2i \end{bmatrix}$. The real and imaginary parts of $\vec{v}e^{2it}$ are

$$\operatorname{Re} \begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{2it} = \begin{bmatrix} \cos(2t) \\ -2\sin(2t) \end{bmatrix}, \quad \operatorname{Im} \begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{2it} = \begin{bmatrix} \sin(2t) \\ 2\cos(2t) \end{bmatrix}.$$

We can take any linear combination of them to get other solutions, which one we take depends on the initial conditions. Now note that the real part is a parametric equation for an ellipse. Same with the imaginary part and in fact any linear combination of the two. This is what happens in general when the eigenvalues are purely imaginary. So when the eigenvalues are purely imaginary, we get *ellipses* for the solutions. This type of picture is sometimes called a *center*. See [Figure 4.9](#).

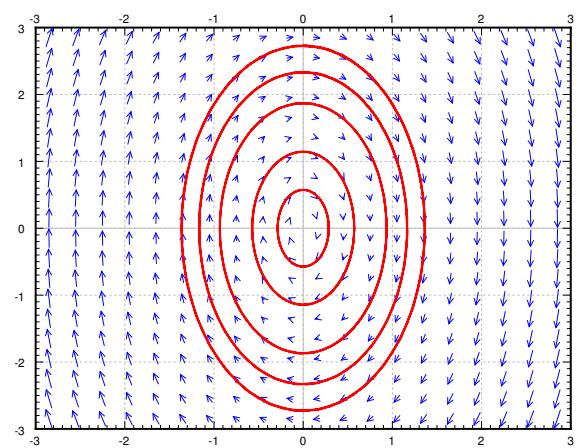


Figure 4.9: Example center vector field.

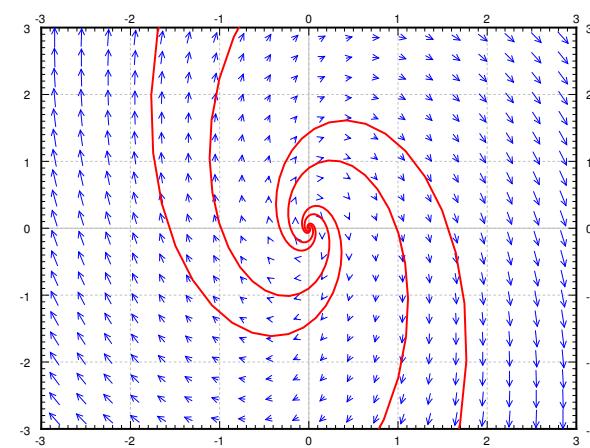


Figure 4.10: Example spiral source vector field.

Case 5. Now suppose the complex eigenvalues have a positive real part. That is, suppose the eigenvalues are $a \pm ib$ for some $a > 0$. For example, let $P = \begin{bmatrix} 1 & 1 \\ -4 & 1 \end{bmatrix}$. The eigenvalues turn out to be $1 \pm 2i$ and eigenvectors are $\begin{bmatrix} 1 \\ 2i \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -2i \end{bmatrix}$. We take $1 + 2i$ and its eigenvector $\begin{bmatrix} 1 \\ 2i \end{bmatrix}$ and find the real and imaginary parts of $\vec{v}e^{(1+2i)t}$ are

$$\operatorname{Re} \begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{(1+2i)t} = e^t \begin{bmatrix} \cos(2t) \\ -2\sin(2t) \end{bmatrix}, \quad \operatorname{Im} \begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{(1+2i)t} = e^t \begin{bmatrix} \sin(2t) \\ 2\cos(2t) \end{bmatrix}.$$

Note the e^t in front of the solutions. The solutions grow in magnitude while spinning around the origin. Hence we get a *spiral source*. See [Figure 4.10](#).

Case 6. Finally suppose the complex eigenvalues have a negative real part. That is, suppose the eigenvalues are $-a \pm ib$ for some $a > 0$. For example, let $P = \begin{bmatrix} -1 & -1 \\ 4 & -1 \end{bmatrix}$. The

eigenvalues turn out to be $-1 \pm 2i$ and eigenvectors are $\begin{bmatrix} 1 \\ -2i \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 2i \end{bmatrix}$. We take $-1 - 2i$ and its eigenvector $\begin{bmatrix} 1 \\ 2i \end{bmatrix}$ and find the real and imaginary parts of $\vec{v}e^{(-1-2i)t}$ are

$$\operatorname{Re} \begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{(-1-2i)t} = e^{-t} \begin{bmatrix} \cos(2t) \\ 2\sin(2t) \end{bmatrix}, \quad \operatorname{Im} \begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{(-1-2i)t} = e^{-t} \begin{bmatrix} -\sin(2t) \\ 2\cos(2t) \end{bmatrix}.$$

Note the e^{-t} in front of the solutions. The solutions shrink in magnitude while spinning around the origin. Hence we get a *spiral sink*. See Figure 4.11.

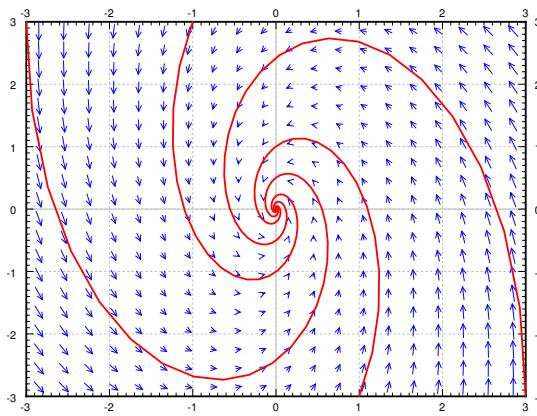


Figure 4.11: Example spiral sink vector field.

Next, we look at what can happen when the eigenvalue in the 2×2 system is repeated. There are a few different options here based on whether there are two linearly independent eigenvectors for that eigenvalue.

Case 7. If we have a repeated eigenvalue with two linearly independent eigenvectors, this means that our matrix A is of the form

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

for the repeated eigenvalue λ . This means that $A\vec{v} = \lambda\vec{v}$ for all vectors \vec{v} . So, every vector is part of a straight line solution, and so every solution goes either directly towards or directly away from the origin. This gives a *proper node* which can be a sink or a source depending on whether the eigenvalue is positive or negative.

Case 8. If we have a repeated eigenvalue with only one linearly independent eigenvector, then we only have one straight-line solution. For instance, the matrix

$$A = \begin{bmatrix} 4 & -1 \\ 1 & 2 \end{bmatrix}$$

has only one eigenvector of $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ for eigenvalue 3. Like the nodal sources and sinks, the solutions will go to zero and infinity along the straight line solutions. In this case, because

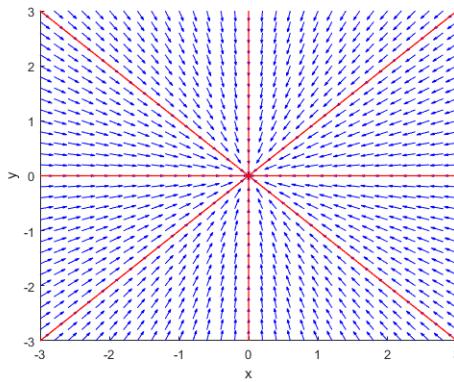


Figure 4.12: Example proper nodal sink vector field.

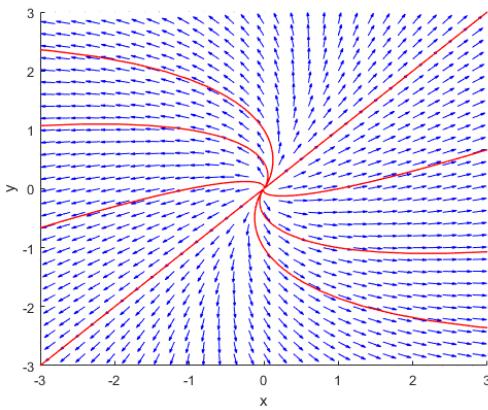


Figure 4.13: Example improper nodal source vector field.

there is only one straight line, the phase portrait looks somewhere between a node and a spiral. This gives an *improper node* which can be a source or sink depending on the sign of the eigenvalue.

We summarize the behavior of linear homogeneous two-dimensional systems given by a nonsingular matrix in [Table 4.1](#). Systems where one of the eigenvalues is zero (the matrix is singular) come up in practice from time to time, see [Example 4.1.2](#) on page 239, and the pictures are somewhat different (simpler in a way). See the exercises.

4.5.1 Exercises

Exercise 4.5.1: Take the equation $mx'' + cx' + kx = 0$, with $m > 0$, $c \geq 0$, $k > 0$ for the mass-spring system.

- Convert this to a system of first order equations.

Eigenvalues	Behavior
real and both positive	source / unstable node
real and both negative	sink / asymptotically stable node
real and opposite signs	saddle
purely imaginary	center point / ellipses
complex with positive real part	spiral source
complex with negative real part	spiral sink
repeated with two eigenvectors	proper node (asympt. stable or unstable)
repeated with one eigenvector	improper node (asympt. stable or unstable)

Table 4.1: Summary of behavior of linear homogeneous two-dimensional systems.

- b) Classify for what m, c, k do you get which behavior.
- c) Can you explain from physical intuition why you do not get all the different kinds of behavior here?

Exercise 4.5.2: What happens in the case when $P = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$? In this case the eigenvalue is repeated and there is only one independent eigenvector. What picture does this look like?

Exercise 4.5.3: What happens in the case when $P = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$? Does this look like any of the pictures we have drawn?

Exercise 4.5.4:* Describe the behavior of the following systems without solving:

- | | |
|---------------------------------------|---|
| a) $x' = x + y, \quad y' = x - y.$ | b) $x'_1 = x_1 + x_2, \quad x'_2 = 2x_2.$ |
| c) $x'_1 = -2x_2, \quad x'_2 = 2x_1.$ | d) $x' = x + 3y, \quad y' = -2x - 4y.$ |
| e) $x' = x - 4y, \quad y' = -4x + y.$ | |

Exercise 4.5.5: Which behaviors are possible if P is diagonal, that is $P = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$? You can assume that a and b are not zero.

Exercise 4.5.6:* Suppose that $\vec{x}' = A\vec{x}$ where A is a 2 by 2 matrix with eigenvalues $2 \pm i$. Describe the behavior.

Exercise 4.5.7:* For each of the following matrices A , describe the behavior and classify the phase portrait of the system given by $\vec{x}' = A\vec{x}$.

- | | |
|---|--|
| a) $A = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix}$ | b) $A = \begin{bmatrix} 3 & 5 \\ -1 & 1 \end{bmatrix}$ |
| c) $A = \begin{bmatrix} 8 & -18 \\ 4 & -10 \end{bmatrix}$ | d) $A = \begin{bmatrix} -2 & -4 \\ 0 & -3 \end{bmatrix}$ |
| e) $A = \begin{bmatrix} 3 & -2 \\ 2 & -3 \end{bmatrix}$ | f) $A = \begin{bmatrix} -3 & -4 \\ 1 & 1 \end{bmatrix}$ |

Exercise 4.5.8: Take the system from [Example 4.1.2](#) on page 239, $x'_1 = \frac{r}{V}(x_2 - x_1)$, $x'_2 = \frac{r}{V}(x_1 - x_2)$. As we said, one of the eigenvalues is zero. What is the other eigenvalue, how does the picture look like and what happens when t goes to infinity.

Exercise 4.5.9:* Take $\begin{bmatrix} x \\ y \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$. Draw the vector field and describe the behavior. Is it one of the behaviors that we have seen before?

4.6 Nonhomogeneous systems

Attribution: [JL], §3.9.

Learning Objectives

After this section, you will be able to:

- Use the eigenvector decomposition or diagonalization to solve non-homogeneous systems,
- Use the method of undetermined coefficients to solve non-homogeneous systems, and
- Use the method of variation of parameters and fundamental matrices of solutions to solve non-homogeneous systems.

Now, we want to take a look at solving non-homogeneous linear systems. As discussed previously, the process here is the same as it was for second order non-homogeneous equations. We can solve the homogeneous equation and then need one particular solution to the non-homogeneous problem. Adding these together gives the general solution to the non-homogeneous problem, where we can pick constants to meet an initial condition if it is given. This section here will focus on a variety of methods to find this particular solution.

4.6.1 First order constant coefficient

Eigenvector decomposition

For this first method, note that eigenvectors of a matrix give the directions in which the matrix acts like a scalar. If we solve the system along these directions, the computations are simpler as we treat the matrix as a scalar. We then put those solutions together to get the general solution for the system.

One way to see how this works is via an example where the eigenvectors are very simple, which happens when the matrix in question is diagonal.

Example 4.6.1: Find the general solution of the non-homogeneous system

$$\vec{x}'(t) = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{2t} \\ e^{-t} \end{bmatrix}.$$

Solution: If we write this system out in components, we get

$$\begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e^{2t} \\ e^{-t} \end{bmatrix},$$

or

$$x'_1 = x_1 + e^{2t} \quad x'_2 = 3x_2 + e^{-t}.$$

These are two completely separated, or *decoupled* equations. We can solve each of these via first-order integrating factor methods. For the first, we get

$$\begin{aligned}x'_1 - x_1 &= e^{2t} \\(e^{-t}x_1)' &= e^t \\e^{-t}x_1 &= e^t + C_1 \\x_1(t) &= e^{2t} + C_1 e^t\end{aligned}$$

and for the second, we see that

$$\begin{aligned}x'_2 - 3x_2 &= e^{-t} \\(e^{-3t}x_2)' &= e^{2t} \\e^{-3t}x_2 &= \frac{1}{2}e^{2t} + C_2 \\x_2(t) &= \frac{1}{2}e^{-t} + C_2 e^{3t}\end{aligned}$$

Therefore, the solution to this system is

$$\begin{bmatrix}x_1 \\ x_2\end{bmatrix} = \begin{bmatrix}e^{2t} + C_1 e^t \\ \frac{1}{2}e^{-t} + C_2 e^{3t}\end{bmatrix}$$

or, rewriting in a different form,

$$\vec{x}(t) = \begin{bmatrix}e^{2t} \\ \frac{1}{2}e^{-t}\end{bmatrix} + C_1 \begin{bmatrix}1 \\ 0\end{bmatrix} e^t + C_2 \begin{bmatrix}0 \\ 1\end{bmatrix} e^{3t}.$$

□

Therefore, if we have a non-linear system with a diagonal matrix, then we can separate the decoupled equations, solve them individually, and put them back together into a full solution. In this particular case, the eigenvectors of A were $\begin{bmatrix}1 \\ 0\end{bmatrix}$ and $\begin{bmatrix}0 \\ 1\end{bmatrix}$, and so the standard basis vectors were the directions in which A acts like a scalar. When the eigenvectors are not the standard basis vectors, we need to take them into account in order to use this method.

Take the equation

$$\vec{x}'(t) = A\vec{x}(t) + \vec{f}(t). \quad (4.4)$$

Assume A has n linearly independent eigenvectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$. Write

$$\vec{x}(t) = \vec{v}_1 \xi_1(t) + \vec{v}_2 \xi_2(t) + \cdots + \vec{v}_n \xi_n(t). \quad (4.5)$$

That is, we wish to write our solution as a linear combination of eigenvectors of A . If we solve for the scalar functions ξ_1 through ξ_n , we have our solution \vec{x} . Let us decompose \vec{f} in terms of the eigenvectors as well. We wish to write

$$\vec{f}(t) = \vec{v}_1 g_1(t) + \vec{v}_2 g_2(t) + \cdots + \vec{v}_n g_n(t). \quad (4.6)$$

That is, we wish to find g_1 through g_n that satisfy (4.6). Since all the eigenvectors are independent, the matrix $E = [\vec{v}_1 \quad \vec{v}_2 \quad \cdots \quad \vec{v}_n]$ is invertible. Write the equation (4.6) as

$\vec{f} = E\vec{g}$, where the components of \vec{g} are the functions g_1 through g_n . Then $\vec{g} = E^{-1}\vec{f}$. Hence it is always possible to find \vec{g} when there are n linearly independent eigenvectors.

We plug (4.5) into (4.4), and note that $A\vec{v}_k = \lambda_k \vec{v}_k$:

$$\begin{aligned} \overbrace{\vec{v}_1\xi'_1 + \vec{v}_2\xi'_2 + \cdots + \vec{v}_n\xi'_n}^{\vec{x}'} &= \overbrace{A(\vec{v}_1\xi_1 + \vec{v}_2\xi_2 + \cdots + \vec{v}_n\xi_n)}^{A\vec{x}} + \overbrace{\vec{v}_1g_1 + \vec{v}_2g_2 + \cdots + \vec{v}_ng_n}^{\vec{f}} \\ &= A\vec{v}_1\xi_1 + A\vec{v}_2\xi_2 + \cdots + A\vec{v}_n\xi_n + \vec{v}_1g_1 + \vec{v}_2g_2 + \cdots + \vec{v}_ng_n \\ &= \vec{v}_1\lambda_1\xi_1 + \vec{v}_2\lambda_2\xi_2 + \cdots + \vec{v}_n\lambda_n\xi_n + \vec{v}_1g_1 + \vec{v}_2g_2 + \cdots + \vec{v}_ng_n \\ &= \vec{v}_1(\lambda_1\xi_1 + g_1) + \vec{v}_2(\lambda_2\xi_2 + g_2) + \cdots + \vec{v}_n(\lambda_n\xi_n + g_n). \end{aligned}$$

If we identify the coefficients of the vectors \vec{v}_1 through \vec{v}_n , we get the equations

$$\begin{aligned} \xi'_1 &= \lambda_1\xi_1 + g_1, \\ \xi'_2 &= \lambda_2\xi_2 + g_2, \\ &\vdots \\ \xi'_n &= \lambda_n\xi_n + g_n. \end{aligned}$$

Each one of these equations is independent of the others. They are all linear first order equations and can easily be solved by the standard integrating factor method for single equations. That is, for the k^{th} equation we write

$$\xi'_k(t) - \lambda_k\xi_k(t) = g_k(t).$$

We use the integrating factor $e^{-\lambda_k t}$ to find that

$$\frac{d}{dt} \left[\xi_k(t) e^{-\lambda_k t} \right] = e^{-\lambda_k t} g_k(t).$$

We integrate and solve for ξ_k to get

$$\xi_k(t) = e^{\lambda_k t} \int e^{-\lambda_k s} g_k(s) ds + C_k e^{\lambda_k t}.$$

If we are looking for just any particular solution, we can set C_k to be zero. If we leave these constants in, we get the general solution. Write $\vec{x}(t) = \vec{v}_1\xi_1(t) + \vec{v}_2\xi_2(t) + \cdots + \vec{v}_n\xi_n(t)$, and we are done.

As always, it is perhaps better to write these integrals as definite integrals. Suppose that we have an initial condition $\vec{x}(0) = \vec{b}$. Take $\vec{a} = E^{-1}\vec{b}$ to find $\vec{b} = \vec{v}_1a_1 + \vec{v}_2a_2 + \cdots + \vec{v}_na_n$, just like before. Then if we write

$$\xi_k(t) = e^{\lambda_k t} \int_0^t e^{-\lambda_k s} g_k(s) ds + a_k e^{\lambda_k t},$$

we get the particular solution $\vec{x}(t) = \vec{v}_1\xi_1(t) + \vec{v}_2\xi_2(t) + \cdots + \vec{v}_n\xi_n(t)$ satisfying $\vec{x}(0) = \vec{b}$, because $\xi_k(0) = a_k$.

Let us remark that the technique we just outlined is the eigenvalue method applied to nonhomogeneous systems. If a system is homogeneous, that is, if $\vec{f} = \vec{0}$, then the equations we get are $\xi'_k = \lambda_k \xi_k$, and so $\xi_k = C_k e^{\lambda_k t}$ are the solutions and that's precisely what we got in § 4.4.

Example 4.6.2: Let $A = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$. Solve $\vec{x}' = A\vec{x} + \vec{f}$ where $\vec{f}(t) = \begin{bmatrix} 2e^t \\ 2t \end{bmatrix}$ for $\vec{x}(0) = \begin{bmatrix} 3/16 \\ -5/16 \end{bmatrix}$.

Solution: The eigenvalues of A are -2 and 4 and corresponding eigenvectors are $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ respectively. This calculation is left as an exercise. We write down the matrix E of the eigenvectors and compute its inverse (using the inverse formula for 2×2 matrices)

$$E = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad E^{-1} = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

We are looking for a solution of the form $\vec{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \xi_1 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \xi_2$. We first need to write \vec{f} in terms of the eigenvectors. That is we wish to write $\vec{f} = \begin{bmatrix} 2e^t \\ 2t \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} g_1 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} g_2$. Thus

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = E^{-1} \begin{bmatrix} 2e^t \\ 2t \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2e^t \\ 2t \end{bmatrix} = \begin{bmatrix} e^t - t \\ e^t + t \end{bmatrix}.$$

So $g_1 = e^t - t$ and $g_2 = e^t + t$.

We further need to write $\vec{x}(0)$ in terms of the eigenvectors. That is, we wish to write $\vec{x}(0) = \begin{bmatrix} 3/16 \\ -5/16 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} a_1 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} a_2$. Hence

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = E^{-1} \begin{bmatrix} 3/16 \\ -5/16 \end{bmatrix} = \begin{bmatrix} 1/4 \\ -1/16 \end{bmatrix}.$$

So $a_1 = 1/4$ and $a_2 = -1/16$. We plug our \vec{x} into the equation and get

$$\begin{aligned} \overbrace{\begin{bmatrix} 1 \\ -1 \end{bmatrix} \xi'_1 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \xi'_2}^{\vec{x}'} &= \overbrace{A \begin{bmatrix} 1 \\ -1 \end{bmatrix} \xi_1 + A \begin{bmatrix} 1 \\ 1 \end{bmatrix} \xi_2}^{A\vec{x}} + \overbrace{\begin{bmatrix} 1 \\ -1 \end{bmatrix} g_1 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} g_2}^{\vec{f}} \\ &= \begin{bmatrix} 1 \\ -1 \end{bmatrix} (-2\xi_1) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} 4\xi_2 + \begin{bmatrix} 1 \\ -1 \end{bmatrix} (e^t - t) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} (e^t + t). \end{aligned}$$

We get the two equations

$$\begin{aligned} \xi'_1 &= -2\xi_1 + e^t - t, & \text{where } \xi_1(0) = a_1 = \frac{1}{4}, \\ \xi'_2 &= 4\xi_2 + e^t + t, & \text{where } \xi_2(0) = a_2 = \frac{-1}{16}. \end{aligned}$$

We solve with integrating factor. Computation of the integral is left as an exercise to the student. You will need integration by parts.

$$\xi_1 = e^{-2t} \int e^{2t} (e^t - t) dt + C_1 e^{-2t} = \frac{e^t}{3} - \frac{t}{2} + \frac{1}{4} + C_1 e^{-2t}.$$

C_1 is the constant of integration. As $\xi_1(0) = 1/4$, then $1/4 = 1/3 + 1/4 + C_1$ and hence $C_1 = -1/3$. Similarly

$$\xi_2 = e^{4t} \int e^{-4t} (e^t + t) dt + C_2 e^{4t} = -\frac{e^t}{3} - \frac{t}{4} - \frac{1}{16} + C_2 e^{4t}.$$

As $\xi_2(0) = -1/16$ we have $-1/16 = -1/3 - 1/16 + C_2$ and hence $C_2 = 1/3$. The solution is

$$\vec{x}(t) = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \underbrace{\left(\frac{e^t - e^{-2t}}{3} + \frac{1 - 2t}{4} \right)}_{\xi_1} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \underbrace{\left(\frac{e^{4t} - e^t}{3} - \frac{4t + 1}{16} \right)}_{\xi_2} = \begin{bmatrix} \frac{e^{4t} - e^{-2t}}{3} + \frac{3 - 12t}{16} \\ \frac{e^{-2t} + e^{4t} - 2e^t}{3} + \frac{4t - 5}{16} \end{bmatrix}.$$

That is, $x_1 = \frac{e^{4t} - e^{-2t}}{3} + \frac{3 - 12t}{16}$ and $x_2 = \frac{e^{-2t} + e^{4t} - 2e^t}{3} + \frac{4t - 5}{16}$. □

Exercise 4.6.1: Check that x_1 and x_2 solve the problem. Check both that they satisfy the differential equation and that they satisfy the initial conditions.

Another way to view this method is via diagonalization of the matrix A . With the matrix E of eigenvectors, and D , a diagonal matrix with the eigenvalues in the same order as the eigenvectors are in E , a fact from linear algebra says that

$$A = EDE^{-1}.$$

If we take the original equation

$$\vec{x}'(t) = A\vec{x}(t) + \vec{f}(t)$$

and define a new vector function \vec{y} by $\vec{x} = E\vec{y}$, plugging this in gives

$$E\vec{y}'(t) = EDE^{-1}E\vec{y}(t) + \vec{f}(t)$$

or

$$\vec{y}'(t) = D\vec{y}(t) + E^{-1}\vec{f}(t)$$

which is now a decoupled system that we can solve by the method in the first example. This also matches the $\vec{\xi}(t)$ solutions found in the second method. The solution can then be converted back to \vec{x} by multiplying by the matrix E .

Undetermined coefficients

The method of undetermined coefficients also works for systems. The only difference is that we use unknown vectors rather than just numbers. Same caveats apply to undetermined coefficients for systems as for single equations. This method does not always work. Furthermore, if the right-hand side is complicated, we have to solve for lots of variables. Each element of an unknown vector is an unknown number. In system of 3 equations with say 4 unknown vectors (this would not be uncommon), we already have 12 unknown numbers to solve for. The method can turn into a lot of tedious work if done by hand. As the method is essentially the same as for single equations, let us just do an example.

Example 4.6.3: Let $A = \begin{bmatrix} -1 & 0 \\ -2 & 1 \end{bmatrix}$. Find a particular solution of $\vec{x}' = A\vec{x} + \vec{f}$ where $\vec{f}(t) = \begin{bmatrix} e^t \\ t \end{bmatrix}$.

Solution: Note that we can solve this system in an easier way (can you see how?), but for the purposes of the example, let us use the eigenvalue method plus undetermined coefficients. The eigenvalues of A are -1 and 1 and corresponding eigenvectors are $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ respectively. Hence our complementary solution is

$$\vec{x}_c = \alpha_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-t} + \alpha_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^t,$$

for some arbitrary constants α_1 and α_2 .

We would want to guess a particular solution of

$$\vec{x} = \vec{a}e^t + \vec{b}t + \vec{c}.$$

However, something of the form $\vec{a}e^t$ appears in the complementary solution. Because we do not yet know if the vector \vec{a} is a multiple of $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, we do not know if a conflict arises. It is possible that there is no conflict, but to be safe we should also try $\vec{b}te^t$. Here we find the crux of the difference between a single equation and systems. We try *both* terms $\vec{a}e^t$ and $\vec{b}te^t$ in the solution, not just the term $\vec{b}te^t$. Therefore, we try

$$\vec{x} = \vec{a}e^t + \vec{b}te^t + \vec{c}t + \vec{d}.$$

Thus we have 8 unknowns. We write $\vec{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$, $\vec{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, $\vec{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$, and $\vec{d} = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}$. We plug \vec{x} into the equation. First let us compute \vec{x}' .

$$\vec{x}' = (\vec{a} + \vec{b})e^t + \vec{b}te^t + \vec{c}t + \vec{d} = \begin{bmatrix} a_1 + b_1 \\ a_2 + b_2 \end{bmatrix} e^t + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} te^t + \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} t + \vec{d}.$$

Now \vec{x}' must equal $A\vec{x} + \vec{f}$, which is

$$\begin{aligned} A\vec{x} + \vec{f} &= A\vec{a}e^t + A\vec{b}te^t + A\vec{c}t + A\vec{d} + \vec{f} \\ &= \begin{bmatrix} -a_1 \\ -2a_1 + a_2 \end{bmatrix} e^t + \begin{bmatrix} -b_1 \\ -2b_1 + b_2 \end{bmatrix} te^t + \begin{bmatrix} -c_1 \\ -2c_1 + c_2 \end{bmatrix} t + \begin{bmatrix} -d_1 \\ -2d_1 + d_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^t + \begin{bmatrix} 0 \\ 1 \end{bmatrix} t \\ &= \begin{bmatrix} -a_1 + 1 \\ -2a_1 + a_2 \end{bmatrix} e^t + \begin{bmatrix} -b_1 \\ -2b_1 + b_2 \end{bmatrix} te^t + \begin{bmatrix} -c_1 \\ -2c_1 + c_2 + 1 \end{bmatrix} t + \begin{bmatrix} -d_1 \\ -2d_1 + d_2 \end{bmatrix}. \end{aligned}$$

We identify the coefficients of e^t , te^t , t and any constant vectors in \vec{x}' and in $A\vec{x} + \vec{f}$ to find the equations:

$$\begin{aligned} a_1 + b_1 &= -a_1 + 1, & 0 &= -c_1, \\ a_2 + b_2 &= -2a_1 + a_2, & 0 &= -2c_1 + c_2 + 1, \\ b_1 &= -b_1, & c_1 &= -d_1, \\ b_2 &= -2b_1 + b_2, & c_2 &= -2d_1 + d_2. \end{aligned}$$

We could write the 8×9 augmented matrix and start row reduction, but it is easier to just solve the equations in an ad hoc manner. Immediately we see that $b_1 = 0$, $c_1 = 0$, $d_1 = 0$. Plugging these back in, we get that $c_2 = -1$ and $d_2 = -1$. The remaining equations that tell us something are

$$\begin{aligned} a_1 &= -a_1 + 1, \\ a_2 + b_2 &= -2a_1 + a_2. \end{aligned}$$

So $a_1 = 1/2$ and $b_2 = -1$. Finally, a_2 can be arbitrary and still satisfy the equations. We are looking for just a single solution so presumably the simplest one is when $a_2 = 0$. Therefore,

$$\vec{x} = \vec{a}e^t + \vec{b}te^t + \vec{c}t + \vec{d} = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix} e^t + \begin{bmatrix} 0 \\ -1 \end{bmatrix} te^t + \begin{bmatrix} 0 \\ -1 \end{bmatrix} t + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}e^t \\ -te^t - t - 1 \end{bmatrix}.$$

That is, $x_1 = \frac{1}{2}e^t$, $x_2 = -te^t - t - 1$. We would add this to the complementary solution to get the general solution of the problem. Notice that both $\vec{a}e^t$ and $\vec{b}te^t$ were really needed. \square

Exercise 4.6.2: Check that x_1 and x_2 solve the problem. Try setting $a_2 = 1$ and check we get a solution as well. What is the difference between the two solutions we obtained (one with $a_2 = 0$ and one with $a_2 = 1$)?

As you can see, other than the handling of conflicts, undetermined coefficients works exactly the same as it did for single equations. However, the computations can get out of hand pretty quickly for systems. The equation we considered was pretty simple.

4.6.2 First order variable coefficient

Variation of parameters

Just as for a single equation, there is the method of variation of parameters. This method works for any linear system, even if it is not constant coefficient, provided we somehow solve the associated homogeneous problem.

Suppose we have the equation

$$\vec{x}' = A(t) \vec{x} + \vec{f}(t). \quad (4.7)$$

Further, suppose we solved the associated homogeneous equation $\vec{x}' = A(t) \vec{x}$ and found a fundamental matrix solution $X(t)$. The general solution to the associated homogeneous equation is $X(t)\vec{c}$ for a constant vector \vec{c} . Just like for variation of parameters for single equation we try the solution to the nonhomogeneous equation of the form

$$\vec{x}_p = X(t) \vec{u}(t),$$

where $\vec{u}(t)$ is a vector-valued function instead of a constant. We substitute \vec{x}_p into (4.7) to obtain

$$\underbrace{X'(t) \vec{u}(t) + X(t) \vec{u}'(t)}_{\vec{x}'_p(t)} = \underbrace{A(t) X(t) \vec{u}(t)}_{A(t) \vec{x}_p(t)} + \vec{f}(t).$$

But $X(t)$ is a fundamental matrix solution to the homogeneous problem. So $X'(t) = A(t)X(t)$, and

$$\underline{X'(t)} \vec{u}(t) + X(t) \vec{u}'(t) = \underline{X'(t)} \vec{u}(t) + \vec{f}(t).$$

Hence $X(t) \vec{u}'(t) = \vec{f}(t)$. If we compute $[X(t)]^{-1}$, then $\vec{u}'(t) = [X(t)]^{-1} \vec{f}(t)$. We integrate to obtain \vec{u} and we have the particular solution $\vec{x}_p = X(t) \vec{u}(t)$. Let us write this as a formula

$$\boxed{\vec{x}_p = X(t) \int [X(t)]^{-1} \vec{f}(t) dt.}$$

Example 4.6.4: Find a particular solution to

$$\vec{x}' = \frac{1}{t^2+1} \begin{bmatrix} t & -1 \\ 1 & t \end{bmatrix} \vec{x} + \begin{bmatrix} t \\ 1 \end{bmatrix} (t^2+1). \quad (4.8)$$

Solution: Here $A = \frac{1}{t^2+1} \begin{bmatrix} t & -1 \\ 1 & t \end{bmatrix}$ is most definitely not constant. Perhaps by a lucky guess, we find that $X = \begin{bmatrix} 1 & -t \\ t & 1 \end{bmatrix}$ solves $X'(t) = A(t)X(t)$. Once we know the complementary solution we can easily find a solution to (4.8). First we find

$$[X(t)]^{-1} = \frac{1}{t^2+1} \begin{bmatrix} 1 & t \\ -t & 1 \end{bmatrix}.$$

Next we know a particular solution to (4.8) is

$$\begin{aligned} \vec{x}_p &= X(t) \int [X(t)]^{-1} \vec{f}(t) dt \\ &= \begin{bmatrix} 1 & -t \\ t & 1 \end{bmatrix} \int \frac{1}{t^2+1} \begin{bmatrix} 1 & t \\ -t & 1 \end{bmatrix} \begin{bmatrix} t \\ 1 \end{bmatrix} (t^2+1) dt \\ &= \begin{bmatrix} 1 & -t \\ t & 1 \end{bmatrix} \int \begin{bmatrix} 2t \\ -t^2+1 \end{bmatrix} dt \\ &= \begin{bmatrix} 1 & -t \\ t & 1 \end{bmatrix} \begin{bmatrix} t^2 \\ -\frac{1}{3}t^3+t \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{3}t^4 \\ \frac{2}{3}t^3+t \end{bmatrix}. \end{aligned}$$

Adding the complementary solution we find the general solution to (4.8):

$$\vec{x} = \begin{bmatrix} 1 & -t \\ t & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \frac{1}{3}t^4 \\ \frac{2}{3}t^3+t \end{bmatrix} = \begin{bmatrix} c_1 - c_2t + \frac{1}{3}t^4 \\ c_2 + (c_1+1)t + \frac{2}{3}t^3 \end{bmatrix}.$$

Exercise 4.6.3: Check that $x_1 = \frac{1}{3}t^4$ and $x_2 = \frac{2}{3}t^3 + t$ really solve (4.8).

In the variation of parameters, we can obtain the general solution by adding in constants of integration. That is, we will add $X(t)\vec{c}$ for a vector of arbitrary constants. But that is precisely the complementary solution.

4.6.3 Exercises

Exercise 4.6.4: Find a particular solution to $x' = x + 2y + 2t$, $y' = 3x + 2y - 4$,

- a) using eigenvector decomposition, b) using undetermined coefficients.

Exercise 4.6.5:* Find a particular solution to $x' = 5x + 4y + t$, $y' = x + 8y - t$,

- a) using eigenvector decomposition, b) using undetermined coefficients.

Exercise 4.6.6: Find the general solution to $x' = 4x + y - 1$, $y' = x + 4y - e^t$,

- a) using eigenvector decomposition, b) using undetermined coefficients.

Exercise 4.6.7:* Find a particular solution to $x' = y + e^t$, $y' = x + e^t$,

- a) using eigenvector decomposition, b) using undetermined coefficients.

Exercise 4.6.8:* Solve $x'_1 = x_2 + t$, $x'_2 = x_1 + t$ with initial conditions $x_1(0) = 1$, $x_2(0) = 2$, using eigenvector decomposition.

Exercise 4.6.9: For each of the following vector functions $\vec{f}(t)$, find the general solution to the system of differential equations given by

$$\vec{x}' = \begin{bmatrix} -1 & -4 \\ 2 & 5 \end{bmatrix} \vec{x} + \vec{f}(t)$$

using any of the methods described in this section. Notice the similarities and differences between using these methods for different non-homogeneous parts.

a) $\vec{f}(t) = \begin{bmatrix} e^{2t} \\ 1 \end{bmatrix}$	b) $\vec{f}(t) = \begin{bmatrix} e^{-t} + 2 \\ e^{4t} - 1 \end{bmatrix}$	c) $\vec{f}(t) = \begin{bmatrix} e^{3t} \\ t \end{bmatrix}$
d) $\vec{f}(t) = \begin{bmatrix} \sin(3t) \\ 1 - \sin(3t) \end{bmatrix}$	e) $\vec{f}(t) = \begin{bmatrix} t + 2 \\ e^{-2t} \end{bmatrix}$	f) $\vec{f}(t) = \begin{bmatrix} te^t \\ 3 \end{bmatrix}$

Exercise 4.6.10: The variation of parameters method can also be applied to constant coefficient systems. Find the general solution of the system

$$\vec{x}' = \begin{bmatrix} 3 & -1 \\ 2 & -2 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{2t} \\ e^t \end{bmatrix}$$

using

- a) eigenvector decomposition b) variation of parameters.

Compare and contrast these methods. You can use undetermined coefficients to check your answer.

Exercise 4.6.11: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} 5 & -6 \\ 3 & -1 \end{bmatrix} \vec{x} + \begin{bmatrix} t \\ 3 \end{bmatrix} \quad \vec{x}(0) = \begin{bmatrix} 1 \\ -3 \end{bmatrix}.$$

Exercise 4.6.12: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} -4 & 2 \\ -9 & 5 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{3t} \\ e^t - 1 \end{bmatrix} \quad \vec{x}(0) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$

Exercise 4.6.13: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} 3 & 2 \\ 0 & 4 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{4t} \\ e^{3t} - t \end{bmatrix} \quad \vec{x}(0) = \begin{bmatrix} 2 \\ -1 \end{bmatrix}.$$

Exercise 4.6.14: Take the equation $\vec{x}' = \begin{bmatrix} \frac{1}{t} & -1 \\ 1 & \frac{1}{t} \end{bmatrix} \vec{x} + \begin{bmatrix} t^2 \\ -t \end{bmatrix}$.

- a) Check that $\vec{x}_c = c_1 \begin{bmatrix} t \sin t \\ -t \cos t \end{bmatrix} + c_2 \begin{bmatrix} t \cos t \\ t \sin t \end{bmatrix}$ is the complementary solution.
- b) Use variation of parameters to find a particular solution.

4.7 Second order systems and applications

Attribution: [JL], §3.6.

Learning Objectives

After this section, you will be able to:

- Use second order systems of differential equations to model physical problems and
- Solve second order systems using diagonalization or eigenvalue methods.

4.7.1 Undamped mass-spring systems

While we did say that we will usually only look at first order systems, it is sometimes more convenient to study the system in the way it arises naturally. For example, suppose we have 3 masses connected by springs between two walls. We could pick any higher number, and the math would be essentially the same, but for simplicity we pick 3 right now. Let us also assume no friction, that is, the system is undamped. The masses are m_1 , m_2 , and m_3 and the spring constants are k_1 , k_2 , k_3 , and k_4 . Let x_1 be the displacement from rest position of the first mass, and x_2 and x_3 the displacement of the second and third mass. We make, as usual, positive values go right (as x_1 grows, the first mass is moving right). See [Figure 4.14](#).

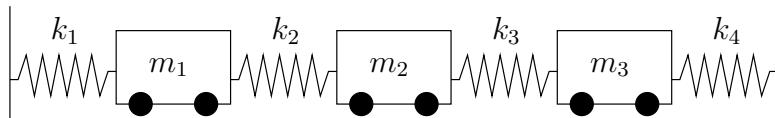


Figure 4.14: System of masses and springs.

This simple system turns up in unexpected places. For example, our world really consists of many small particles of matter interacting together. When we try the system above with many more masses, we obtain a good approximation to how an elastic material behaves.

Let us set up the equations for the three mass system. By Hooke's law, the force acting on the mass equals the spring compression times the spring constant. By Newton's second law, force is mass times acceleration. So if we sum the forces acting on each mass, put the right sign in front of each term, depending on the direction in which it is acting, and set this equal to mass times the acceleration, we end up with the desired system of equations.

$$\begin{aligned} m_1 x_1'' &= -k_1 x_1 + k_2(x_2 - x_1) & = -(k_1 + k_2)x_1 + k_2 x_2, \\ m_2 x_2'' &= -k_2(x_2 - x_1) + k_3(x_3 - x_2) & = k_2 x_1 - (k_2 + k_3)x_2 + k_3 x_3, \\ m_3 x_3'' &= -k_3(x_3 - x_2) - k_4 x_3 & = k_3 x_2 - (k_3 + k_4)x_3. \end{aligned}$$

We define the matrices

$$M = \begin{bmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix} \quad \text{and} \quad K = \begin{bmatrix} -(k_1 + k_2) & k_2 & 0 \\ k_2 & -(k_2 + k_3) & k_3 \\ 0 & k_3 & -(k_3 + k_4) \end{bmatrix}.$$

We write the equation simply as

$$M\vec{x}'' = K\vec{x}.$$

At this point we could introduce 3 new variables and write out a system of 6 first order equations. We claim this simple setup is easier to handle as a second order system. We call \vec{x} the *displacement vector*, M the *mass matrix*, and K the *stiffness matrix*.

Exercise 4.7.1: Repeat this setup for 4 masses (find the matrices M and K). Do it for 5 masses. Can you find a prescription to do it for n masses?

As with a single equation we want to “divide by M . ” This means computing the inverse of M . The masses are all nonzero and M is a diagonal matrix, so computing the inverse is easy:

$$M^{-1} = \begin{bmatrix} \frac{1}{m_1} & 0 & 0 \\ 0 & \frac{1}{m_2} & 0 \\ 0 & 0 & \frac{1}{m_3} \end{bmatrix}.$$

This fact follows readily by how we multiply diagonal matrices. As an exercise, you should verify that $MM^{-1} = M^{-1}M = I$.

Let $A = M^{-1}K$. We look at the system $\vec{x}'' = M^{-1}K\vec{x}$, or

$$\vec{x}'' = A\vec{x}.$$

Many real world systems can be modeled by this equation. For simplicity, we will only talk about the given masses-and-springs problem. We try a solution of the form

$$\vec{x} = \vec{v}e^{\alpha t}.$$

We compute that for this guess, $\vec{x}'' = \alpha^2\vec{v}e^{\alpha t}$. We plug our guess into the equation and get

$$\alpha^2\vec{v}e^{\alpha t} = A\vec{v}e^{\alpha t}.$$

We divide by $e^{\alpha t}$ to arrive at $\alpha^2\vec{v} = A\vec{v}$. Hence if α^2 is an eigenvalue of A and \vec{v} is a corresponding eigenvector, we have found a solution.

In our example, and in other common applications, A has only real negative eigenvalues (and possibly a zero eigenvalue). So we study only this case. When an eigenvalue λ is negative, it means that $\alpha^2 = \lambda$ is negative. Hence there is some real number ω such that $-\omega^2 = \lambda$. Then $\alpha = \pm i\omega$. The solution we guessed was

$$\vec{x} = \vec{v}(\cos(\omega t) + i \sin(\omega t)).$$

By taking the real and imaginary parts (note that \vec{v} is real), we find that $\vec{v}\cos(\omega t)$ and $\vec{v}\sin(\omega t)$ are linearly independent solutions.

If an eigenvalue is zero, it turns out that both \vec{v} and $\vec{v}t$ are solutions, where \vec{v} is an eigenvector corresponding to the eigenvalue 0.

Exercise 4.7.2: Show that if A has a zero eigenvalue and \vec{v} is a corresponding eigenvector, then $\vec{x} = \vec{v}(a + bt)$ is a solution of $\vec{x}'' = A\vec{x}$ for arbitrary constants a and b .

Theorem 4.7.1

Let A be a real $n \times n$ matrix with n distinct real negative (or zero) eigenvalues we denote by $-\omega_1^2 > -\omega_2^2 > \dots > -\omega_n^2$, and corresponding eigenvectors by $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$. If A is invertible (that is, if $\omega_1 > 0$), then

$$\vec{x}(t) = \sum_{i=1}^n \vec{v}_i (a_i \cos(\omega_i t) + b_i \sin(\omega_i t)),$$

is the general solution of

$$\vec{x}'' = A\vec{x},$$

for some arbitrary constants a_i and b_i . If A has a zero eigenvalue, that is $\omega_1 = 0$, and all other eigenvalues are distinct and negative, then the general solution can be written as

$$\vec{x}(t) = \vec{v}_1(a_1 + b_1 t) + \sum_{i=2}^n \vec{v}_i (a_i \cos(\omega_i t) + b_i \sin(\omega_i t)).$$

We use this solution and the setup from the introduction of this section even when some of the masses and springs are missing. For example, when there are only 2 masses and only 2 springs, simply take only the equations for the two masses and set all the spring constants for the springs that are missing to zero.

4.7.2 Examples

Example 4.7.1: Consider the setup in Figure 4.15, with $m_1 = 2 \text{ kg}$, $m_2 = 1 \text{ kg}$, $k_1 = 4 \text{ N/m}$, and $k_2 = 2 \text{ N/m}$.

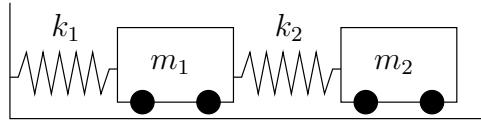


Figure 4.15: System of masses and springs.

Solution: The equations we write down are

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \vec{x}'' = \begin{bmatrix} -(4+2) & 2 \\ 2 & -2 \end{bmatrix} \vec{x},$$

or

$$\vec{x}'' = \begin{bmatrix} -3 & 1 \\ 2 & -2 \end{bmatrix} \vec{x}.$$

We find the eigenvalues of A to be $\lambda = -1, -4$ (exercise). We find corresponding eigenvectors to be $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ respectively (exercise).

We check the theorem and note that $\omega_1 = 1$ and $\omega_2 = 2$. Hence the general solution is

$$\vec{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} (a_1 \cos(t) + b_1 \sin(t)) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} (a_2 \cos(2t) + b_2 \sin(2t)).$$

The two terms in the solution represent the two so-called *natural* or *normal modes of oscillation*. And the two (angular) frequencies are the *natural frequencies*. The first natural frequency is 1, and second natural frequency is 2. The two modes are plotted in Figure 4.16.

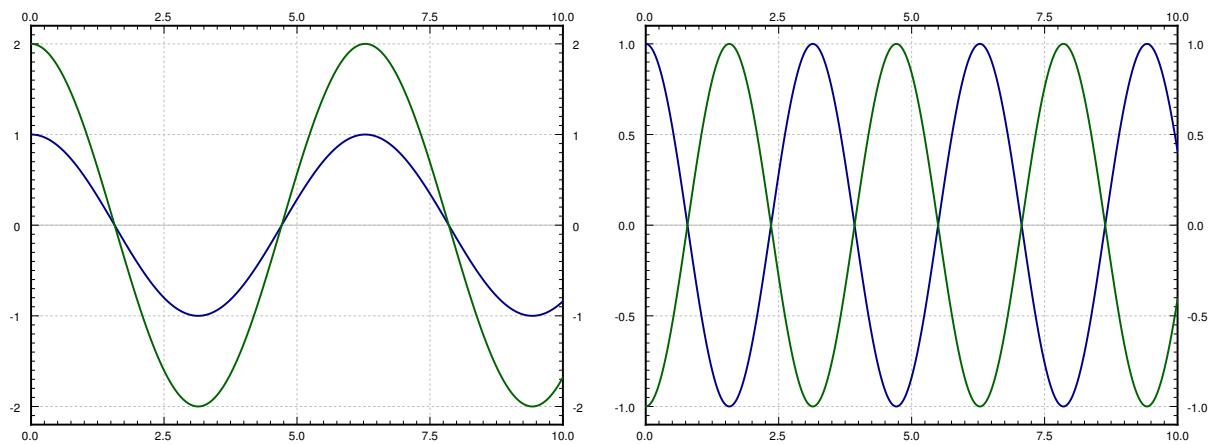


Figure 4.16: The two modes of the mass-spring system. In the left plot the masses are moving in unison and in the right plot are masses moving in the opposite direction.

Let us write the solution as

$$\vec{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} c_1 \cos(t - \alpha_1) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} c_2 \cos(2t - \alpha_2).$$

The first term,

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} c_1 \cos(t - \alpha_1) = \begin{bmatrix} c_1 \cos(t - \alpha_1) \\ 2c_1 \cos(t - \alpha_1) \end{bmatrix},$$

corresponds to the mode where the masses move synchronously in the same direction.

The second term,

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} c_2 \cos(2t - \alpha_2) = \begin{bmatrix} c_2 \cos(2t - \alpha_2) \\ -c_2 \cos(2t - \alpha_2) \end{bmatrix},$$

corresponds to the mode where the masses move synchronously but in opposite directions.

The general solution is a combination of the two modes. That is, the initial conditions determine the amplitude and phase shift of each mode. As an example, suppose we have initial conditions

$$\vec{x}(0) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \vec{x}'(0) = \begin{bmatrix} 0 \\ 6 \end{bmatrix}.$$

We use the a_j, b_j constants to solve for initial conditions. First

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \vec{x}(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} a_1 + \begin{bmatrix} 1 \\ -1 \end{bmatrix} a_2 = \begin{bmatrix} a_1 + a_2 \\ 2a_1 - a_2 \end{bmatrix}.$$

We solve (exercise) to find $a_1 = 0, a_2 = 1$. To find the b_1 and b_2 , we differentiate first:

$$\vec{x}' = \begin{bmatrix} 1 \\ 2 \end{bmatrix} (-a_1 \sin(t) + b_1 \cos(t)) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} (-2a_2 \sin(2t) + 2b_2 \cos(2t)).$$

Now we solve:

$$\begin{bmatrix} 0 \\ 6 \end{bmatrix} = \vec{x}'(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} b_1 + \begin{bmatrix} 1 \\ -1 \end{bmatrix} 2b_2 = \begin{bmatrix} b_1 + 2b_2 \\ 2b_1 - 2b_2 \end{bmatrix}.$$

Again solve (exercise) to find $b_1 = 2, b_2 = -1$. So our solution is

$$\vec{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} 2 \sin(t) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} (\cos(2t) - \sin(2t)) = \begin{bmatrix} 2 \sin(t) + \cos(2t) - \sin(2t) \\ 4 \sin(t) - \cos(2t) + \sin(2t) \end{bmatrix}.$$

The graphs of the two displacements, x_1 and x_2 of the two carts is in [Figure 4.17](#).

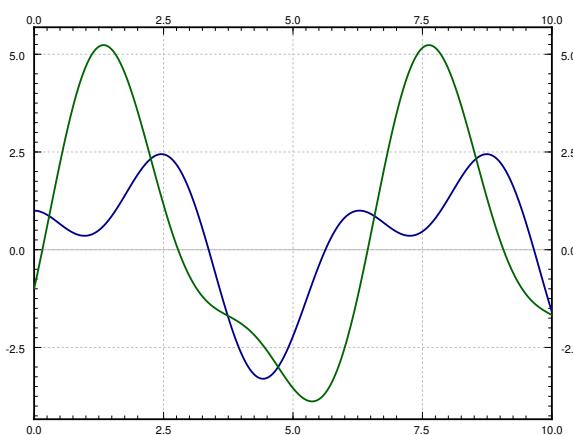


Figure 4.17: Superposition of the two modes given the initial conditions.

Example 4.7.2: We have two toy rail cars. Car 1 of mass 2 kg is traveling at 3 m/s towards the second rail car of mass 1 kg. There is a bumper on the second rail car that engages at the moment the cars hit (it connects to two cars) and does not let go. The bumper acts like a spring of spring constant $k = 2 \text{ N/m}$. The second car is 10 meters from a wall. See [Figure 4.18](#) on the following page.

We want to ask several questions. At what time after the cars link does impact with the wall happen? What is the speed of car 2 when it hits the wall?

Solution: OK, let us first set the system up. Let $t = 0$ be the time when the two cars link up. Let x_1 be the displacement of the first car from the position at $t = 0$, and let x_2 be the displacement of the second car from its original location. Then the time when $x_2(t) = 10$ is

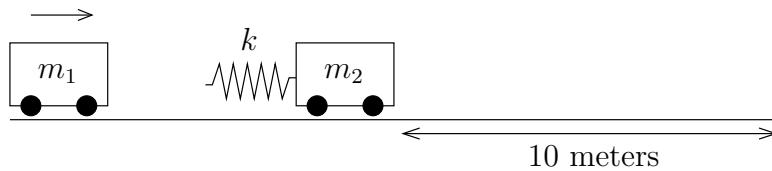


Figure 4.18: The crash of two rail cars.

exactly the time when impact with wall occurs. For this t , $x'_2(t)$ is the speed at impact. This system acts just like the system of the previous example but without k_1 . Hence the equation is

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \vec{x}'' = \begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \vec{x},$$

or

$$\vec{x}'' = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix} \vec{x}.$$

We compute the eigenvalues of A . It is not hard to see that the eigenvalues are 0 and -3 (exercise). Furthermore, eigenvectors are $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -2 \end{bmatrix}$ respectively (exercise). Then $\omega_1 = 0$, $\omega_2 = \sqrt{3}$, and by the second part of the theorem the general solution is

$$\begin{aligned} \vec{x} &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} (a_1 + b_1 t) + \begin{bmatrix} 1 \\ -2 \end{bmatrix} (a_2 \cos(\sqrt{3}t) + b_2 \sin(\sqrt{3}t)) \\ &= \begin{bmatrix} a_1 + b_1 t + a_2 \cos(\sqrt{3}t) + b_2 \sin(\sqrt{3}t) \\ a_1 + b_1 t - 2a_2 \cos(\sqrt{3}t) - 2b_2 \sin(\sqrt{3}t) \end{bmatrix}. \end{aligned}$$

We now apply the initial conditions. First the cars start at position 0 so $x_1(0) = 0$ and $x_2(0) = 0$. The first car is traveling at 3 m/s, so $x'_1(0) = 3$ and the second car starts at rest, so $x'_2(0) = 0$. The first condition says

$$\vec{0} = \vec{x}(0) = \begin{bmatrix} a_1 + a_2 \\ a_1 - 2a_2 \end{bmatrix}.$$

It is not hard to see that $a_1 = a_2 = 0$. We set $a_1 = 0$ and $a_2 = 0$ in $\vec{x}(t)$ and differentiate to get

$$\vec{x}'(t) = \begin{bmatrix} b_1 + \sqrt{3}b_2 \cos(\sqrt{3}t) \\ b_1 - 2\sqrt{3}b_2 \cos(\sqrt{3}t) \end{bmatrix}.$$

So

$$\begin{bmatrix} 3 \\ 0 \end{bmatrix} = \vec{x}'(0) = \begin{bmatrix} b_1 + \sqrt{3}b_2 \\ b_1 - 2\sqrt{3}b_2 \end{bmatrix}.$$

Solving these two equations we find $b_1 = 2$ and $b_2 = \frac{1}{\sqrt{3}}$. Hence the position of our cars is (until the impact with the wall)

$$\vec{x} = \begin{bmatrix} 2t + \frac{1}{\sqrt{3}} \sin(\sqrt{3}t) \\ 2t - \frac{2}{\sqrt{3}} \sin(\sqrt{3}t) \end{bmatrix}.$$

Note how the presence of the zero eigenvalue resulted in a term containing t . This means that the cars will be traveling in the positive direction as time grows, which is what we expect.

What we are really interested in is the second expression, the one for x_2 . We have $x_2(t) = 2t - \frac{2}{\sqrt{3}} \sin(\sqrt{3}t)$. See [Figure 4.19](#) for the plot of x_2 versus time.

Just from the graph we can see that time of impact will be a little more than 5 seconds from time zero. For this we have to solve the equation $10 = x_2(t) = 2t - \frac{2}{\sqrt{3}} \sin(\sqrt{3}t)$. Using a computer (or even a graphing calculator) we find that $t_{\text{impact}} \approx 5.22$ seconds.

The speed of the second car is $x'_2 = 2 - 2 \cos(\sqrt{3}t)$. At the time of impact (5.22 seconds from $t = 0$) we get $x'_2(t_{\text{impact}}) \approx 3.85$. The maximum speed is the maximum of $2 - 2 \cos(\sqrt{3}t)$, which is 4. We are traveling at almost the maximum speed when we hit the wall.

Suppose that Bob is a tiny person sitting on car 2. Bob has a Martini in his hand and would like not to spill it. Let us suppose Bob would not spill his Martini when the first car links up with car 2, but if car 2 hits the wall at any speed greater than zero, Bob will spill his drink. Suppose Bob can move car 2 a few meters towards or away from the wall (he cannot go all the way to the wall, nor can he get out of the way of the first car). Is there a “safe” distance for him to be at? A distance such that the impact with the wall is at zero speed?

The answer is yes. Looking at [Figure 4.19](#), we note the “plateau” between $t = 3$ and $t = 4$. There is a point where the speed is zero. To find it we solve $x'_2(t) = 0$. This is when $\cos(\sqrt{3}t) = 1$ or in other words when $t = \frac{2\pi}{\sqrt{3}}, \frac{4\pi}{\sqrt{3}}, \dots$ and so on. We plug in the first value to obtain $x_2\left(\frac{2\pi}{\sqrt{3}}\right) = \frac{4\pi}{\sqrt{3}} \approx 7.26$. So a “safe” distance is about 7 and a quarter meters from the wall.

Alternatively Bob could move away from the wall towards the incoming car 2, where another safe distance is $x_2\left(\frac{4\pi}{\sqrt{3}}\right) = \frac{8\pi}{\sqrt{3}} \approx 14.51$ and so on. We can use all the different t such that $x'_2(t) = 0$. Of course $t = 0$ is also a solution, corresponding to $x_2 = 0$, but that means standing right at the wall. □

4.7.3 Forced oscillations

Finally we move to forced oscillations. Suppose that now our system is

$$\vec{x}'' = A\vec{x} + \vec{F} \cos(\omega t). \quad (4.9)$$

That is, we are adding periodic forcing to the system in the direction of the vector \vec{F} .

As before, this system just requires us to find one particular solution \vec{x}_p , add it to the general solution of the associated homogeneous system \vec{x}_c , and we will have the general

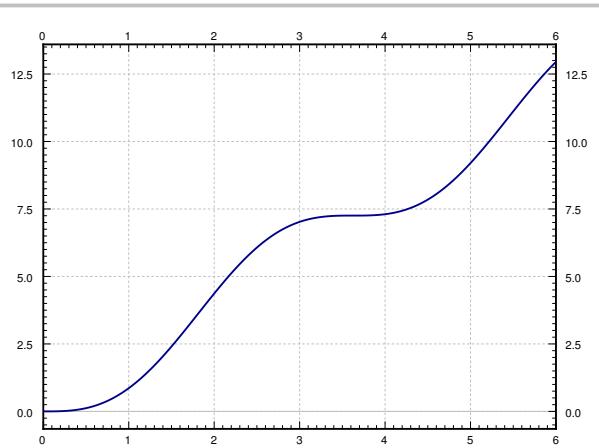


Figure 4.19: Position of the second car in time (ignoring the wall).

solution to (4.9). Let us suppose that ω is not one of the natural frequencies of $\vec{x}'' = A\vec{x}$, then we can guess

$$\vec{x}_p = \vec{c}\cos(\omega t),$$

where \vec{c} is an unknown constant vector. Note that we do not need to use sine since there are only second derivatives. We solve for \vec{c} to find \vec{x}_p . This is really just the method of *undetermined coefficients* for systems. Let us differentiate \vec{x}_p twice to get

$$\vec{x}_p'' = -\omega^2\vec{c}\cos(\omega t).$$

Plug \vec{x}_p and \vec{x}_p'' into equation (4.9):

$$\underbrace{-\omega^2\vec{c}\cos(\omega t)}_{\vec{x}_p''} = \underbrace{A\vec{c}\cos(\omega t)}_{A\vec{x}_p} + \vec{F}\cos(\omega t).$$

We cancel out the cosine and rearrange the equation to obtain

$$(A + \omega^2 I)\vec{c} = -\vec{F}.$$

So

$$\vec{c} = (A + \omega^2 I)^{-1}(-\vec{F}).$$

Of course this is possible only if $(A + \omega^2 I) = (A - (-\omega^2)I)$ is invertible. That matrix is invertible if and only if $-\omega^2$ is not an eigenvalue of A . That is true if and only if ω is not a natural frequency of the system.

We simplified things a little bit. If we wish to have the forcing term to be in the units of force, say Newtons, then we must write

$$M\vec{x}'' = K\vec{x} + \vec{G}\cos(\omega t).$$

If we then write things in terms of $A = M^{-1}K$, we have

$$\vec{x}'' = M^{-1}K\vec{x} + M^{-1}\vec{G}\cos(\omega t) \quad \text{or} \quad \vec{x}'' = A\vec{x} + \vec{F}\cos(\omega t),$$

where $\vec{F} = M^{-1}\vec{G}$.

Example 4.7.3: Let us take the example in Figure 4.15 on page 293 with the same parameters as before: $m_1 = 2$, $m_2 = 1$, $k_1 = 4$, and $k_2 = 2$. Now suppose that there is a force $2\cos(3t)$ acting on the second cart.

Solution: The equation is

$$\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \vec{x}'' = \begin{bmatrix} -4 & 2 \\ 2 & -2 \end{bmatrix} \vec{x} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} \cos(3t) \quad \text{or} \quad \vec{x}'' = \begin{bmatrix} -3 & 1 \\ 2 & -2 \end{bmatrix} \vec{x} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} \cos(3t).$$

We solved the associated homogeneous equation before and found the complementary solution to be

$$\vec{x}_c = \begin{bmatrix} 1 \\ 2 \end{bmatrix} (a_1 \cos(t) + b_1 \sin(t)) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} (a_2 \cos(2t) + b_2 \sin(2t)).$$

The natural frequencies are 1 and 2. As 3 is not a natural frequency, we try $\vec{c}\cos(3t)$. We invert $(A + 3^2 I)$:

$$\left(\begin{bmatrix} -3 & 1 \\ 2 & -2 \end{bmatrix} + 3^2 I \right)^{-1} = \begin{bmatrix} 6 & 1 \\ 2 & 7 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{7}{40} & \frac{-1}{40} \\ \frac{-1}{20} & \frac{3}{20} \end{bmatrix}.$$

Hence,

$$\vec{c} = (A + \omega^2 I)^{-1}(-\vec{F}) = \begin{bmatrix} \frac{7}{40} & \frac{-1}{40} \\ \frac{-1}{20} & \frac{3}{20} \end{bmatrix} \begin{bmatrix} 0 \\ -2 \end{bmatrix} = \begin{bmatrix} \frac{1}{20} \\ \frac{-3}{10} \end{bmatrix}.$$

Combining with the general solution of the associated homogeneous problem, we get that the general solution to $\vec{x}'' = A\vec{x} + \vec{F}\cos(\omega t)$ is

$$\vec{x} = \vec{x}_c + \vec{x}_p = \begin{bmatrix} 1 \\ 2 \end{bmatrix} (a_1 \cos(t) + b_1 \sin(t)) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} (a_2 \cos(2t) + b_2 \sin(2t)) + \begin{bmatrix} \frac{1}{20} \\ \frac{-3}{10} \end{bmatrix} \cos(3t).$$

We then solve for the constants a_1 , a_2 , b_1 , and b_2 using any initial conditions we are given.]

Note that given force \vec{f} , we write the equation as $M\vec{x}'' = K\vec{x} + \vec{f}$ to get the units right. Then we write $\vec{x}'' = M^{-1}K\vec{x} + M^{-1}\vec{f}$. The term $\vec{g} = M^{-1}\vec{f}$ in $\vec{x}'' = A\vec{x} + \vec{g}$ is in units of force per unit mass.

If ω is a natural frequency of the system, *resonance* may occur, because we will have to try a particular solution of the form

$$\vec{x}_p = \vec{c}t \sin(\omega t) + \vec{d} \cos(\omega t).$$

That is assuming that the eigenvalues of the coefficient matrix are distinct. Next, note that the amplitude of this solution grows without bound as t grows.

4.7.4 Non-Homogeneous Solutions

Undetermined coefficients

We have already seen a simple example of the method of undetermined coefficients for second order systems in § 4.7. This method is essentially the same as undetermined coefficients for first order systems. There are some simplifications that we can make, as we did in § 4.7. Let the equation be

$$\vec{x}'' = A\vec{x} + \vec{F}(t),$$

where A is a constant matrix. If $\vec{F}(t)$ is of the form $\vec{F}_0 \cos(\omega t)$, then as two derivatives of cosine is again cosine we can try a solution of the form

$$\vec{x}_p = \vec{c}\cos(\omega t),$$

and we do not need to introduce sines.

If the \vec{F} is a sum of cosines, note that we still have the superposition principle. If $\vec{F}(t) = \vec{F}_0 \cos(\omega_0 t) + \vec{F}_1 \cos(\omega_1 t)$, then we would try $\vec{a}\cos(\omega_0 t)$ for the problem $\vec{x}'' = A\vec{x} + \vec{F}_0 \cos(\omega_0 t)$,

and we would try $\vec{b} \cos(\omega_1 t)$ for the problem $\vec{x}'' = A\vec{x} + \vec{F}_1 \cos(\omega_1 t)$. Then we sum the solutions.

However, if there is duplication with the complementary solution, or the equation is of the form $\vec{x}'' = A\vec{x}' + B\vec{x} + \vec{F}(t)$, then we need to do the same thing as we do for first order systems.

You will never go wrong with putting in more terms than needed into your guess. You will find that the extra coefficients will turn out to be zero. But it is useful to save some time and effort.

Eigenvector decomposition

If we have the system

$$\vec{x}'' = A\vec{x} + \vec{f}(t),$$

we can do *eigenvector decomposition*, just like for first order systems.

Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the eigenvalues and $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$ be eigenvectors. Again form the matrix $E = [\vec{v}_1 \quad \vec{v}_2 \quad \dots \quad \vec{v}_n]$. Write

$$\vec{x}(t) = \vec{v}_1 \xi_1(t) + \vec{v}_2 \xi_2(t) + \dots + \vec{v}_n \xi_n(t).$$

Decompose \vec{f} in terms of the eigenvectors

$$\vec{f}(t) = \vec{v}_1 g_1(t) + \vec{v}_2 g_2(t) + \dots + \vec{v}_n g_n(t),$$

where, again, $\vec{g} = E^{-1}\vec{f}$.

We plug in, and as before we obtain

$$\begin{aligned} \overbrace{\vec{v}_1 \xi_1'' + \vec{v}_2 \xi_2'' + \dots + \vec{v}_n \xi_n''}^{\vec{x}''} &= \overbrace{A(\vec{v}_1 \xi_1 + \vec{v}_2 \xi_2 + \dots + \vec{v}_n \xi_n)}^{A\vec{x}} + \overbrace{\vec{v}_1 g_1 + \vec{v}_2 g_2 + \dots + \vec{v}_n g_n}^{\vec{f}} \\ &= A\vec{v}_1 \xi_1 + A\vec{v}_2 \xi_2 + \dots + A\vec{v}_n \xi_n + \vec{v}_1 g_1 + \vec{v}_2 g_2 + \dots + \vec{v}_n g_n \\ &= \vec{v}_1 \lambda_1 \xi_1 + \vec{v}_2 \lambda_2 \xi_2 + \dots + \vec{v}_n \lambda_n \xi_n + \vec{v}_1 g_1 + \vec{v}_2 g_2 + \dots + \vec{v}_n g_n \\ &= \vec{v}_1 (\lambda_1 \xi_1 + g_1) + \vec{v}_2 (\lambda_2 \xi_2 + g_2) + \dots + \vec{v}_n (\lambda_n \xi_n + g_n). \end{aligned}$$

We identify the coefficients of the eigenvectors to get the equations

$$\xi_1'' = \lambda_1 \xi_1 + g_1,$$

$$\xi_2'' = \lambda_2 \xi_2 + g_2,$$

⋮

$$\xi_n'' = \lambda_n \xi_n + g_n.$$

Each one of these equations is independent of the others. We solve each equation using the methods of [chapter 2](#). We write $\vec{x}(t) = \vec{v}_1 \xi_1(t) + \vec{v}_2 \xi_2(t) + \dots + \vec{v}_n \xi_n(t)$, and we are done; we have a particular solution. We find the general solutions for ξ_1 through ξ_n , and again $\vec{x}(t) = \vec{v}_1 \xi_1(t) + \vec{v}_2 \xi_2(t) + \dots + \vec{v}_n \xi_n(t)$ is the general solution (and not just a particular solution).

Example 4.7.4: Let us do the same example from before using this method.

Solution: The equation is

$$\vec{x}'' = \begin{bmatrix} -3 & 1 \\ 2 & -2 \end{bmatrix} \vec{x} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} \cos(3t).$$

The eigenvalues are -1 and -4 , with eigenvectors $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Therefore $E = \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}$ and $E^{-1} = \frac{1}{3} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}$. Therefore,

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = E^{-1} \vec{f}(t) = \frac{1}{3} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \cos(3t) \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \cos(3t) \\ \frac{-2}{3} \cos(3t) \end{bmatrix}.$$

So after the whole song and dance of plugging in, the equations we get are

$$\xi_1'' = -\xi_1 + \frac{2}{3} \cos(3t), \quad \xi_2'' = -4\xi_2 - \frac{2}{3} \cos(3t).$$

For each equation we use the method of undetermined coefficients. We try $C_1 \cos(3t)$ for the first equation and $C_2 \cos(3t)$ for the second equation. We plug in to get

$$\begin{aligned} -9C_1 \cos(3t) &= -C_1 \cos(3t) + \frac{2}{3} \cos(3t), \\ -9C_2 \cos(3t) &= -4C_2 \cos(3t) - \frac{2}{3} \cos(3t). \end{aligned}$$

We solve each of these equations separately. We get $-9C_1 = -C_1 + 2/3$ and $-9C_2 = -4C_2 - 2/3$. And hence $C_1 = -1/12$ and $C_2 = 2/15$. So our particular solution is

$$\vec{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \left(\frac{-1}{12} \cos(3t) \right) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} \left(\frac{2}{15} \cos(3t) \right) = \begin{bmatrix} 1/20 \\ -3/10 \end{bmatrix} \cos(3t).$$

This solution matches what we got previously. □

4.7.5 Exercises

Exercise 4.7.3: Find a particular solution to

$$\vec{x}'' = \begin{bmatrix} -3 & 1 \\ 2 & -2 \end{bmatrix} \vec{x} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} \cos(2t).$$

Exercise 4.7.4:* Find the general solution to $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \vec{x}'' = \begin{bmatrix} -3 & 0 & 0 \\ 2 & -4 & 0 \\ 0 & 6 & -3 \end{bmatrix} \vec{x} + \begin{bmatrix} \cos(2t) \\ 0 \\ 0 \end{bmatrix}$.

Exercise 4.7.5 (challenging): Let us take the example in [Figure 4.15](#) on page 293 with the same parameters as before: $m_1 = 2$, $k_1 = 4$, and $k_2 = 2$, except for m_2 , which is unknown. Suppose that there is a force $\cos(5t)$ acting on the first mass. Find an m_2 such that there exists a particular solution where the first mass does not move.

Note: This idea is called dynamic damping. In practice there will be a small amount of damping and so any transient solution will disappear and after long enough time, the first mass will always come to a stop.

Exercise 4.7.6: Let us take the *Example 4.7.2* on page 295, but that at time of impact, car 2 is moving to the left at the speed of 3 m/s.

- a) Find the behavior of the system after linkup.
- b) Will the second car hit the wall, or will it be moving away from the wall as time goes on?
- c) At what speed would the first car have to be traveling for the system to essentially stay in place after linkup?

Exercise 4.7.7: Let us take the example in *Figure 4.15* on page 293 with parameters $m_1 = m_2 = 1$, $k_1 = k_2 = 1$. Does there exist a set of initial conditions for which the first cart moves but the second cart does not? If so, find those conditions. If not, argue why not.

Exercise 4.7.8:* Suppose there are three carts of equal mass m and connected by two springs of constant k (and no connections to walls). Set up the system and find its general solution.

Exercise 4.7.9:* Suppose a cart of mass 2 kg is attached by a spring of constant $k = 1$ to a cart of mass 3 kg, which is attached to the wall by a spring also of constant $k = 1$. Suppose that the initial position of the first cart is 1 meter in the positive direction from the rest position, and the second mass starts at the rest position. The masses are not moving and are let go. Find the position of the second mass as a function of time.

Exercise 4.7.10: Find the general solution to $x_1'' = -6x_1 + 3x_2 + \cos(t)$, $x_2'' = 2x_1 - 7x_2 + 3\cos(t)$,

- a) using eigenvector decomposition,
- b) using undetermined coefficients.

Exercise 4.7.11: Find the general solution to $x_1'' = -6x_1 + 3x_2 + \cos(2t)$, $x_2'' = 2x_1 - 7x_2 + 3\cos(2t)$,

- a) using eigenvector decomposition,
- b) using undetermined coefficients.

Exercise 4.7.12:* Solve $x_1'' = -3x_1 + x_2 + t$, $x_2'' = 9x_1 + 5x_2 + \cos(t)$ with initial conditions $x_1(0) = 0$, $x_2(0) = 0$, $x_1'(0) = 0$, $x_2'(0) = 0$, using eigenvector decomposition.

4.8 Matrix exponentials

Attribution: [JL], §3.8.

Learning Objectives

After this section, you will be able to:

- Compute the exponential of a matrix and
- Use the matrix exponential to solve linear systems of differential equations.

4.8.1 Definition

There is another way of finding a fundamental matrix solution of a system. Consider the constant coefficient equation

$$\vec{x}' = P\vec{x}.$$

If this would be just one equation (when P is a number or a 1×1 matrix), then the solution would be

$$\vec{x} = e^{Pt}.$$

That doesn't make sense if P is a larger matrix, but essentially the same computation that led to the above works for matrices when we define e^{Pt} properly. First let us write down the Taylor series for e^{at} for some number a :

$$e^{at} = 1 + at + \frac{(at)^2}{2} + \frac{(at)^3}{6} + \frac{(at)^4}{24} + \cdots = \sum_{k=0}^{\infty} \frac{(at)^k}{k!}.$$

Recall $k! = 1 \cdot 2 \cdot 3 \cdots k$ is the factorial, and $0! = 1$. We differentiate this series term by term

$$\frac{d}{dt} (e^{at}) = 0 + a + a^2t + \frac{a^3t^2}{2} + \frac{a^4t^3}{6} + \cdots = a \left(1 + at + \frac{(at)^2}{2} + \frac{(at)^3}{6} + \cdots \right) = ae^{at}.$$

Maybe we can try the same trick with matrices.

Definition 4.8.1

For an $n \times n$ matrix A we define the *matrix exponential* as

$$e^A \stackrel{\text{def}}{=} I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 + \cdots + \frac{1}{k!}A^k + \cdots$$

Let us not worry about convergence. The series really does always converge. We usually write Pt as tP by convention when P is a matrix. With this small change and by the exact same calculation as above we have that

$$\frac{d}{dt} (e^{tP}) = Pe^{tP}.$$

Now P and hence e^{tP} is an $n \times n$ matrix. What we are looking for is a vector. In the 1×1 case we would at this point multiply by an arbitrary constant to get the general solution. In the matrix case we multiply by a column vector \vec{c} .

Theorem 4.8.1

Let P be an $n \times n$ matrix. Then the general solution to $\vec{x}' = P\vec{x}$ is

$$\vec{x} = e^{tP}\vec{c},$$

where \vec{c} is an arbitrary constant vector. In fact, $\vec{x}(0) = \vec{c}$.

Let us check:

$$\frac{d}{dt}\vec{x} = \frac{d}{dt}(e^{tP}\vec{c}) = Pe^{tP}\vec{c} = P\vec{x}.$$

Hence e^{tP} is a fundamental matrix solution of the homogeneous system. So if we can compute the matrix exponential, we have another method of solving constant coefficient homogeneous systems. It also makes it easy to solve for initial conditions. To solve $\vec{x}' = A\vec{x}$, $\vec{x}(0) = \vec{b}$, we take the solution

$$\vec{x} = e^{tA}\vec{b}.$$

This equation follows because $e^{0A} = I$, so $\vec{x}(0) = e^{0A}\vec{b} = \vec{b}$.

We mention a drawback of matrix exponentials. In general $e^{A+B} \neq e^A e^B$. The trouble is that matrices do not commute, that is, in general $AB \neq BA$. If you try to prove $e^{A+B} \neq e^A e^B$ using the Taylor series, you will see why the lack of commutativity becomes a problem. However, it is still true that if $AB = BA$, that is, if A and B commute, then $e^{A+B} = e^A e^B$. We will find this fact useful. Let us restate this as a theorem to make a point.

Theorem 4.8.2

If $AB = BA$, then $e^{A+B} = e^A e^B$. Otherwise, $e^{A+B} \neq e^A e^B$ in general.

4.8.2 Simple cases

In some instances it may work to just plug into the series definition. Suppose the matrix is diagonal. For example, $D = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$. Then

$$D^k = \begin{bmatrix} a^k & 0 \\ 0 & b^k \end{bmatrix},$$

and

$$\begin{aligned} e^D &= I + D + \frac{1}{2}D^2 + \frac{1}{6}D^3 + \dots \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} + \frac{1}{2} \begin{bmatrix} a^2 & 0 \\ 0 & b^2 \end{bmatrix} + \frac{1}{6} \begin{bmatrix} a^3 & 0 \\ 0 & b^3 \end{bmatrix} + \dots = \begin{bmatrix} e^a & 0 \\ 0 & e^b \end{bmatrix}. \end{aligned}$$

So by this rationale

$$e^I = \begin{bmatrix} e & 0 \\ 0 & e \end{bmatrix} \quad \text{and} \quad e^{aI} = \begin{bmatrix} e^a & 0 \\ 0 & e^a \end{bmatrix}.$$

This makes exponentials of certain other matrices easy to compute. For example, the matrix $A = \begin{bmatrix} 5 & 4 \\ -1 & 1 \end{bmatrix}$ can be written as $3I + B$ where $B = \begin{bmatrix} 2 & 4 \\ -1 & 2 \end{bmatrix}$. Notice that $B^2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$. So $B^k = 0$ for all $k \geq 2$. Therefore, $e^B = I + B$. Suppose we actually want to compute e^{tA} . The matrices $3tI$ and tB commute (exercise: check this) and $e^{tB} = I + tB$, since $(tB)^2 = t^2 B^2 = 0$. We write

$$\begin{aligned} e^{tA} &= e^{3tI+tB} = e^{3tI}e^{tB} = \begin{bmatrix} e^{3t} & 0 \\ 0 & e^{3t} \end{bmatrix} (I + tB) = \\ &= \begin{bmatrix} e^{3t} & 0 \\ 0 & e^{3t} \end{bmatrix} \begin{bmatrix} 1+2t & 4t \\ -t & 1-2t \end{bmatrix} = \begin{bmatrix} (1+2t)e^{3t} & 4te^{3t} \\ -te^{3t} & (1-2t)e^{3t} \end{bmatrix}. \end{aligned}$$

We found a fundamental matrix solution for the system $\vec{x}' = A\vec{x}$. Note that this matrix has a repeated eigenvalue with a defect; there is only one eigenvector for the eigenvalue 3. So we found a perhaps easier way to handle this case. In fact, if a matrix A is 2×2 and has an eigenvalue λ of multiplicity 2, then either $A = \lambda I$, or $A = \lambda I + B$ where $B^2 = 0$. This is a good exercise.

Exercise 4.8.1: Suppose that A is 2×2 and λ is the only eigenvalue. Show that $(A - \lambda I)^2 = 0$, and therefore that we can write $A = \lambda I + B$, where $B^2 = 0$ (and possibly $B = 0$). Hint: First write down what does it mean for the eigenvalue to be of multiplicity 2. You will get an equation for the entries. Now compute the square of B .

Matrices B such that $B^k = 0$ for some k are called *nilpotent*. Computation of the matrix exponential for nilpotent matrices is easy by just writing down the first k terms of the Taylor series.

4.8.3 General matrices

In general, the exponential is not as easy to compute as above. We usually cannot write a matrix as a sum of commuting matrices where the exponential is simple for each one. But fear not, it is still not too difficult provided we can find enough eigenvectors. First we need the following interesting result about matrix exponentials. For two square matrices A and B , with B invertible, we have

$$e^{BAB^{-1}} = Be^A B^{-1}.$$

This can be seen by writing down the Taylor series. First

$$(BAB^{-1})^2 = BAB^{-1}BAB^{-1} = BAIAB^{-1} = BA^2B^{-1}.$$

And by the same reasoning $(BAB^{-1})^k = BAB^{-1}$. Now write the Taylor series for $e^{BAB^{-1}}$:

$$\begin{aligned} e^{BAB^{-1}} &= I + BAB^{-1} + \frac{1}{2}(BAB^{-1})^2 + \frac{1}{6}(BAB^{-1})^3 + \dots \\ &= BB^{-1} + BAB^{-1} + \frac{1}{2}BA^2B^{-1} + \frac{1}{6}BA^3B^{-1} + \dots \\ &= B(I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 + \dots)B^{-1} \\ &= Be^A B^{-1}. \end{aligned}$$

Given a square matrix A , we can usually write $A = EDE^{-1}$, where D is diagonal and E invertible. This procedure is called *diagonalization*. If we can do that, the computation of the exponential becomes easy as e^D is just taking the exponential of the entries on the diagonal. Adding t into the mix, we can then compute the exponential

$$e^{tA} = Ee^{tD}E^{-1}.$$

To diagonalize A we need n linearly independent eigenvectors of A . Otherwise, this method of computing the exponential does not work and we need to be trickier, but we will not get into such details. Let E be the matrix with the eigenvectors as columns. Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be the eigenvalues and let $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$ be the eigenvectors, then $E = [\vec{v}_1 \quad \vec{v}_2 \quad \dots \quad \vec{v}_n]$. Make a diagonal matrix D with the eigenvalues on the diagonal:

$$D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

We compute

$$\begin{aligned} AE &= A[\vec{v}_1 \quad \vec{v}_2 \quad \dots \quad \vec{v}_n] \\ &= [A\vec{v}_1 \quad A\vec{v}_2 \quad \dots \quad A\vec{v}_n] \\ &= [\lambda_1\vec{v}_1 \quad \lambda_2\vec{v}_2 \quad \dots \quad \lambda_n\vec{v}_n] \\ &= [\vec{v}_1 \quad \vec{v}_2 \quad \dots \quad \vec{v}_n]D \\ &= ED. \end{aligned}$$

The columns of E are linearly independent as these are linearly independent eigenvectors of A . Hence E is invertible. Since $AE = ED$, we multiply on the right by E^{-1} and we get

$$A = EDE^{-1}.$$

This means that $e^A = Ee^D E^{-1}$. Multiplying the matrix by t we obtain

$$e^{tA} = Ee^{tD}E^{-1} = E \begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n t} \end{bmatrix} E^{-1}. \quad (4.10)$$

The formula (4.10), therefore, gives the formula for computing a fundamental matrix solution e^{tA} for the system $\vec{x}' = A\vec{x}$, in the case where we have n linearly independent eigenvectors.

This computation still works when the eigenvalues and eigenvectors are complex, though then you have to compute with complex numbers. It is clear from the definition that if A is real, then e^{tA} is real. So you will only need complex numbers in the computation and not for the result. You may need to apply Euler's formula to simplify the result. If simplified properly, the final matrix will not have any complex numbers in it.

Example 4.8.1: Compute a fundamental matrix solution using the matrix exponential for the system

$$\begin{bmatrix} x \\ y \end{bmatrix}' = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

Then compute the particular solution for the initial conditions $x(0) = 4$ and $y(0) = 2$.

Let A be the coefficient matrix $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$. We first compute (exercise) that the eigenvalues are 3 and -1 and corresponding eigenvectors are $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Hence the diagonalization of A is

$$\underbrace{\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}}_A = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_E \underbrace{\begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}}_D \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}}_{E^{-1}}^{-1}.$$

We write

$$\begin{aligned} e^{tA} &= Ee^{tD}E^{-1} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} e^{3t} & 0 \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} e^{3t} & 0 \\ 0 & e^{-t} \end{bmatrix} \frac{-1}{2} \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix} \\ &= \frac{-1}{2} \begin{bmatrix} e^{3t} & e^{-t} \\ e^{3t} & -e^{-t} \end{bmatrix} \begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix} \\ &= \frac{-1}{2} \begin{bmatrix} -e^{3t} - e^{-t} & -e^{3t} + e^{-t} \\ -e^{3t} + e^{-t} & -e^{3t} - e^{-t} \end{bmatrix} = \begin{bmatrix} \frac{e^{3t} + e^{-t}}{2} & \frac{e^{3t} - e^{-t}}{2} \\ \frac{e^{3t} - e^{-t}}{2} & \frac{e^{3t} + e^{-t}}{2} \end{bmatrix}. \end{aligned}$$

The initial conditions are $x(0) = 4$ and $y(0) = 2$. Hence, by the property that $e^{0A} = I$ we find that the particular solution we are looking for is $e^{tA}\vec{b}$ where \vec{b} is $\begin{bmatrix} 4 \\ 2 \end{bmatrix}$. Then the particular solution we are looking for is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \frac{e^{3t} + e^{-t}}{2} & \frac{e^{3t} - e^{-t}}{2} \\ \frac{e^{3t} - e^{-t}}{2} & \frac{e^{3t} + e^{-t}}{2} \end{bmatrix} \begin{bmatrix} 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 2e^{3t} + 2e^{-t} + e^{3t} - e^{-t} \\ 2e^{3t} - 2e^{-t} + e^{3t} + e^{-t} \end{bmatrix} = \begin{bmatrix} 3e^{3t} + e^{-t} \\ 3e^{3t} - e^{-t} \end{bmatrix}.$$

4.8.4 Fundamental matrix solutions

We note that if you can compute a fundamental matrix solution in a different way, you can use this to find the matrix exponential e^{tA} . A fundamental matrix solution of a system of ODEs is not unique. The exponential is the fundamental matrix solution with the property

that for $t = 0$ we get the identity matrix. So we must find the right fundamental matrix solution. Let X be any fundamental matrix solution to $\vec{x}' = A\vec{x}$. Then we claim

$$e^{tA} = X(t) [X(0)]^{-1}.$$

Clearly, if we plug $t = 0$ into $X(t) [X(0)]^{-1}$ we get the identity. We can multiply a fundamental matrix solution on the right by any constant invertible matrix and we still get a fundamental matrix solution. All we are doing is changing what are the arbitrary constants in the general solution $\vec{x}(t) = X(t) \vec{c}$.

4.8.5 Approximations

If you think about it, the computation of any fundamental matrix solution X using the eigenvalue method is just as difficult as the computation of e^{tA} . So perhaps we did not gain much by this new tool. However, the Taylor series expansion actually gives us a way to approximate solutions, which the eigenvalue method did not.

The simplest thing we can do is to just compute the series up to a certain number of terms. There are better ways to approximate the exponential*. In many cases however, few terms of the Taylor series give a reasonable approximation for the exponential and may suffice for the application. For example, let us compute the first 4 terms of the series for the matrix $A = [\frac{1}{2} \frac{2}{1}]$.

$$\begin{aligned} e^{tA} &\approx I + tA + \frac{t^2}{2}A^2 + \frac{t^3}{6}A^3 = I + t \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} + t^2 \begin{bmatrix} \frac{5}{2} & 2 \\ 2 & \frac{5}{2} \end{bmatrix} + t^3 \begin{bmatrix} \frac{13}{6} & \frac{7}{3} \\ \frac{7}{3} & \frac{13}{6} \end{bmatrix} = \\ &= \begin{bmatrix} 1 + t + \frac{5}{2}t^2 + \frac{13}{6}t^3 & 2t + 2t^2 + \frac{7}{3}t^3 \\ 2t + 2t^2 + \frac{7}{3}t^3 & 1 + t + \frac{5}{2}t^2 + \frac{13}{6}t^3 \end{bmatrix}. \end{aligned}$$

Just like the scalar version of the Taylor series approximation, the approximation will be better for small t and worse for larger t . For larger t , we will generally have to compute more terms. Let us see how we stack up against the real solution with $t = 0.1$. The approximate solution is approximately (rounded to 8 decimal places)

$$e^{0.1A} \approx I + 0.1A + \frac{0.1^2}{2}A^2 + \frac{0.1^3}{6}A^3 = \begin{bmatrix} 1.12716667 & 0.22233333 \\ 0.22233333 & 1.12716667 \end{bmatrix}.$$

And plugging $t = 0.1$ into the real solution (rounded to 8 decimal places) we get

$$e^{0.1A} = \begin{bmatrix} 1.12734811 & 0.22251069 \\ 0.22251069 & 1.12734811 \end{bmatrix}.$$

Not bad at all! Although if we take the same approximation for $t = 1$ we get

$$I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 = \begin{bmatrix} 6.66666667 & 6.33333333 \\ 6.33333333 & 6.66666667 \end{bmatrix},$$

*C. Moler and C.F. Van Loan, *Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later*, SIAM Review 45 (1), 2003, 3–49

while the real value is (again rounded to 8 decimal places)

$$e^A = \begin{bmatrix} 10.22670818 & 9.85882874 \\ 9.85882874 & 10.22670818 \end{bmatrix}.$$

So the approximation is not very good once we get up to $t = 1$. To get a good approximation at $t = 1$ (say up to 2 decimal places) we would need to go up to the 11th power (exercise).

4.8.6 Non-Homogeneous Systems

Integrating factor

Let us first focus on the nonhomogeneous first order equation

$$\vec{x}'(t) = A\vec{x}(t) + \vec{f}(t),$$

where A is a constant matrix. The method we look at here is the *integrating factor method*. For simplicity we rewrite the equation as

$$\vec{x}'(t) + P\vec{x}(t) = \vec{f}(t),$$

where $P = -A$. We multiply both sides of the equation by e^{tP} (being mindful that we are dealing with matrices that may not commute) to obtain

$$e^{tP}\vec{x}'(t) + e^{tP}P\vec{x}(t) = e^{tP}\vec{f}(t).$$

We notice that $Pe^{tP} = e^{tP}P$. This fact follows by writing down the series definition of e^{tP} :

$$\begin{aligned} Pe^{tP} &= P \left(I + tP + \frac{1}{2}(tP)^2 + \dots \right) = P + tP^2 + \frac{1}{2}t^2P^3 + \dots = \\ &= \left(I + tP + \frac{1}{2}(tP)^2 + \dots \right) P = e^{tP}P. \end{aligned}$$

So $\frac{d}{dt}(e^{tP}) = Pe^{tP} = e^{tP}P$. The product rule says

$$\frac{d}{dt}(e^{tP}\vec{x}(t)) = e^{tP}\vec{x}'(t) + e^{tP}P\vec{x}(t),$$

and so

$$\frac{d}{dt}(e^{tP}\vec{x}(t)) = e^{tP}\vec{f}(t).$$

We can now integrate. That is, we integrate each component of the vector separately

$$e^{tP}\vec{x}(t) = \int e^{tP}\vec{f}(t) dt + \vec{c}.$$

Recall from [Exercise 4.8.10](#) that $(e^{tP})^{-1} = e^{-tP}$. Therefore, we obtain

$$\vec{x}(t) = e^{-tP} \int e^{tP}\vec{f}(t) dt + e^{-tP}\vec{c}.$$

Perhaps it is better understood as a definite integral. In this case it will be easy to also solve for the initial conditions. Consider the equation with initial conditions

$$\vec{x}'(t) + P\vec{x}(t) = \vec{f}(t), \quad \vec{x}(0) = \vec{b}.$$

The solution can then be written as

$$\boxed{\vec{x}(t) = e^{-tP} \int_0^t e^{sP} \vec{f}(s) ds + e^{-tP} \vec{b}.} \quad (4.11)$$

Again, the integration means that each component of the vector $e^{sP} \vec{f}(s)$ is integrated separately. It is not hard to see that (4.11) really does satisfy the initial condition $\vec{x}(0) = \vec{b}$.

$$\vec{x}(0) = e^{-0P} \int_0^0 e^{sP} \vec{f}(s) ds + e^{-0P} \vec{b} = I\vec{b} = \vec{b}.$$

Example 4.8.2: Suppose that we have the system

$$\begin{aligned} x'_1 + 5x_1 - 3x_2 &= e^t, \\ x'_2 + 3x_1 - x_2 &= 0, \end{aligned}$$

with initial conditions $x_1(0) = 1, x_2(0) = 0$.

Solution: Let us write the system as

$$\vec{x}' + \begin{bmatrix} 5 & -3 \\ 3 & -1 \end{bmatrix} \vec{x} = \begin{bmatrix} e^t \\ 0 \end{bmatrix}, \quad \vec{x}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The matrix $P = \begin{bmatrix} 5 & -3 \\ 3 & -1 \end{bmatrix}$ has a doubled eigenvalue 2 with defect 1, and we leave it as an exercise to double check we computed e^{tP} correctly. Once we have e^{tP} , we find e^{-tP} , simply by negating t .

$$e^{tP} = \begin{bmatrix} (1+3t)e^{2t} & -3te^{2t} \\ 3te^{2t} & (1-3t)e^{2t} \end{bmatrix}, \quad e^{-tP} = \begin{bmatrix} (1-3t)e^{-2t} & 3te^{-2t} \\ -3te^{-2t} & (1+3t)e^{-2t} \end{bmatrix}.$$

Instead of computing the whole formula at once, let us do it in stages. First

$$\begin{aligned} \int_0^t e^{sP} \vec{f}(s) ds &= \int_0^t \begin{bmatrix} (1+3s)e^{2s} & -3se^{2s} \\ 3se^{2s} & (1-3s)e^{2s} \end{bmatrix} \begin{bmatrix} e^s \\ 0 \end{bmatrix} ds \\ &= \int_0^t \begin{bmatrix} (1+3s)e^{3s} \\ 3se^{3s} \end{bmatrix} ds \\ &= \begin{bmatrix} \int_0^t (1+3s)e^{3s} ds \\ \int_0^t 3se^{3s} ds \end{bmatrix} \\ &= \begin{bmatrix} te^{3t} \\ \frac{(3t-1)e^{3t}+1}{3} \end{bmatrix} \quad (\text{used integration by parts}). \end{aligned}$$

Then

$$\begin{aligned}
 \vec{x}(t) &= e^{-tP} \int_0^t e^{sP} \vec{f}(s) ds + e^{-tP} \vec{b} \\
 &= \begin{bmatrix} (1-3t)e^{-2t} & 3te^{-2t} \\ -3te^{-2t} & (1+3t)e^{-2t} \end{bmatrix} \begin{bmatrix} te^{3t} \\ \frac{(3t-1)e^{3t}+1}{3} \end{bmatrix} + \begin{bmatrix} (1-3t)e^{-2t} & 3te^{-2t} \\ -3te^{-2t} & (1+3t)e^{-2t} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
 &= \begin{bmatrix} te^{-2t} \\ -\frac{e^t}{3} + \left(\frac{1}{3} + t\right)e^{-2t} \end{bmatrix} + \begin{bmatrix} (1-3t)e^{-2t} \\ -3te^{-2t} \end{bmatrix} \\
 &= \begin{bmatrix} (1-2t)e^{-2t} \\ -\frac{e^t}{3} + \left(\frac{1}{3} - 2t\right)e^{-2t} \end{bmatrix}.
 \end{aligned}$$

Phew!

Let us check that this really works.

$$x'_1 + 5x_1 - 3x_2 = (4te^{-2t} - 4e^{-2t}) + 5(1-2t)e^{-2t} + e^t - (1-6t)e^{-2t} = e^t.$$

Similarly (exercise) $x'_2 + 3x_1 - x_2 = 0$. The initial conditions are also satisfied (exercise). \square

For systems, the integrating factor method only works if P does not depend on t , that is, P is constant. The problem is that in general

$$\frac{d}{dt} \left[e^{\int P(t) dt} \right] \neq P(t) e^{\int P(t) dt},$$

because matrix multiplication is not commutative.

4.8.7 Exercises

Exercise 4.8.2: Using the matrix exponential, find a fundamental matrix solution for the system $x' = 3x + y$, $y' = x + 3y$.

Exercise 4.8.3: Find e^{tA} for the matrix $A = \begin{bmatrix} 2 & 3 \\ 0 & 2 \end{bmatrix}$.

Exercise 4.8.4:* Compute e^{tA} where $A = \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix}$.

Exercise 4.8.5:* Compute e^{tA} where $A = \begin{bmatrix} 1 & -3 & 2 \\ -2 & 1 & 2 \\ -1 & -3 & 4 \end{bmatrix}$.

Exercise 4.8.6:*

- a) Compute e^{tA} where $A = \begin{bmatrix} 3 & -1 \\ 1 & 1 \end{bmatrix}$. b) Solve $\vec{x}' = A\vec{x}$ for $\vec{x}(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Exercise 4.8.7: Find a fundamental matrix solution for the system $x'_1 = 7x_1 + 4x_2 + 12x_3$, $x'_2 = x_1 + 2x_2 + x_3$, $x'_3 = -3x_1 - 2x_2 - 5x_3$. Then find the solution that satisfies $\vec{x}(0) = \begin{bmatrix} 0 \\ 1 \\ -2 \end{bmatrix}$.

Exercise 4.8.8: Compute the matrix exponential e^A for $A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$.

Exercise 4.8.9 (challenging): Suppose $AB = BA$. Show that under this assumption, $e^{A+B} = e^A e^B$.

Exercise 4.8.10: Use [Exercise 4.8.9](#) to show that $(e^A)^{-1} = e^{-A}$. In particular this means that e^A is invertible even if A is not.

Exercise 4.8.11: Let A be a 2×2 matrix with eigenvalues -1 , 1 , and corresponding eigenvectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

- a) Find matrix A with these properties.
- b) Find a fundamental matrix solution to $\vec{x}' = A\vec{x}$.
- c) Solve the system in with initial conditions $\vec{x}(0) = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$.

Exercise 4.8.12: Suppose that A is an $n \times n$ matrix with a repeated eigenvalue λ of multiplicity n . Suppose that there are n linearly independent eigenvectors. Show that the matrix is diagonal, in particular $A = \lambda I$. Hint: Use diagonalization and the fact that the identity matrix commutes with every other matrix.

Exercise 4.8.13: Let $A = \begin{bmatrix} -1 & -1 \\ 1 & -3 \end{bmatrix}$.

- a) Find e^{tA} .
- b) Solve $\vec{x}' = A\vec{x}$, $\vec{x}(0) = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$.

Exercise 4.8.14: Let $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$. Approximate e^{tA} by expanding the power series up to the third order.

Exercise 4.8.15:* Compute the first 3 terms (up to the second degree) of the Taylor expansion of e^{tA} where $A = \begin{bmatrix} 2 & 3 \\ 2 & 2 \end{bmatrix}$ (Write as a single matrix). Then use it to approximate $e^{0.1A}$.

Exercise 4.8.16: For any positive integer n , find a formula (or a recipe) for A^n for the following matrices:

$$\text{a)} \begin{bmatrix} 3 & 0 \\ 0 & 9 \end{bmatrix} \quad \text{b)} \begin{bmatrix} 5 & 2 \\ 4 & 7 \end{bmatrix} \quad \text{c)} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{d)} \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$$

Exercise 4.8.17:* For any positive integer n , find a formula (or a recipe) for A^n for the following matrices:

$$\text{a)} \begin{bmatrix} 7 & 4 \\ -5 & -2 \end{bmatrix} \quad \text{b)} \begin{bmatrix} -3 & 4 \\ -6 & -7 \end{bmatrix} \quad \text{c)} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Exercise 4.8.18: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} 5 & -6 \\ 3 & -1 \end{bmatrix} \vec{x} + \begin{bmatrix} t \\ 3 \end{bmatrix} \quad \vec{x}(0) = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

using matrix exponentials.

Exercise 4.8.19: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} -4 & 2 \\ -9 & 5 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{3t} \\ e^t - 1 \end{bmatrix} \quad \vec{x}(0) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

using matrix exponentials.

Exercise 4.8.20: Solve the initial value problem

$$\vec{x}' = \begin{bmatrix} 3 & 2 \\ 0 & 4 \end{bmatrix} \vec{x} + \begin{bmatrix} e^{4t} \\ e^{3t} - t \end{bmatrix} \quad \vec{x}(0) = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

using matrix exponentials.

Chapter 5

Nonlinear systems

5.1 Linearization, critical points, and equilibria

Attribution: [JL], §8.1.

Learning Objectives

After this section, you will be able to:

- Find critical points of a non-linear system of differential equations and
- Linearize a non-linear system around a critical point by computing the Jacobian matrix.

Except for a few brief detours in chapter 1, we considered mostly linear equations. Linear equations suffice in many applications, but in reality most phenomena require nonlinear equations. Nonlinear equations, however, are notoriously more difficult to understand than linear ones, and many strange new phenomena appear when we allow our equations to be nonlinear.

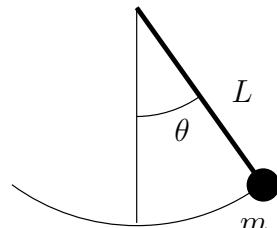
Not to worry, we did not waste all this time studying linear equations. Nonlinear equations can often be approximated by linear ones if we only need a solution “locally,” for example, only for a short period of time, or only for certain parameters. Understanding linear equations can also give us qualitative understanding about a more general nonlinear problem. The idea is similar to what you did in calculus in trying to approximate a function by a line with the right slope.

In § 2.4 we looked at the pendulum of length L . The goal was to solve for the angle $\theta(t)$ as a function of the time t . The equation for the setup is the nonlinear equation

$$\theta'' + \frac{g}{L} \sin \theta = 0.$$

Instead of solving this equation, we solved the rather easier linear equation

$$\theta'' + \frac{g}{L} \theta = 0.$$



While the solution to the linear equation is not exactly what we were looking for, it is rather close to the original, as long as the angle θ is small and the time period involved is short.

You might ask: Why don't we just solve the nonlinear problem? Well, it might be very difficult, impractical, or impossible to solve analytically, depending on the equation in question. We may not even be interested in the actual solution, we might only be interested in some qualitative idea of what the solution is doing. For example, what happens as time goes to infinity?

5.1.1 Autonomous systems and phase plane analysis

We restrict our attention to a two-dimensional autonomous system

$$x' = f(x, y), \quad y' = g(x, y),$$

where $f(x, y)$ and $g(x, y)$ are functions of two variables, and the derivatives are taken with respect to time t . Solutions are functions $x(t)$ and $y(t)$ such that

$$x'(t) = f(x(t), y(t)), \quad y'(t) = g(x(t), y(t)).$$

The way we will analyze the system is very similar to § 1.7, where we studied a single autonomous equation. The ideas in two dimensions are the same, but the behavior can be far more complicated.

It may be best to think of the system of equations as the single vector equation

$$\begin{bmatrix} x \\ y \end{bmatrix}' = \begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix}. \quad (5.1)$$

As in § 4.1 we draw the *phase portrait* (or *phase diagram*), where each point (x, y) corresponds to a specific state of the system. We draw the *vector field* given at each point (x, y) by the vector $\begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix}$. And as before if we find solutions, we draw the trajectories by plotting all points $(x(t), y(t))$ for a certain range of t .

Example 5.1.1: Consider the second order equation $x'' = -x + x^2$. Write this equation as a first order nonlinear system

$$x' = y, \quad y' = -x + x^2.$$

The phase portrait with some trajectories is drawn in Figure 5.1 on the next page.

From the phase portrait it should be clear that even this simple system has fairly complicated behavior. Some trajectories keep oscillating around the origin, and some go off towards infinity. We will return to this example often, and analyze it completely in this (and the next) section.

If we zoom into the diagram near a point where $\begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix}$ is not zero, then nearby the arrows point generally in essentially that same direction and have essentially the same magnitude. In other words the behavior is not that interesting near such a point. We are of course assuming that $f(x, y)$ and $g(x, y)$ are continuous.

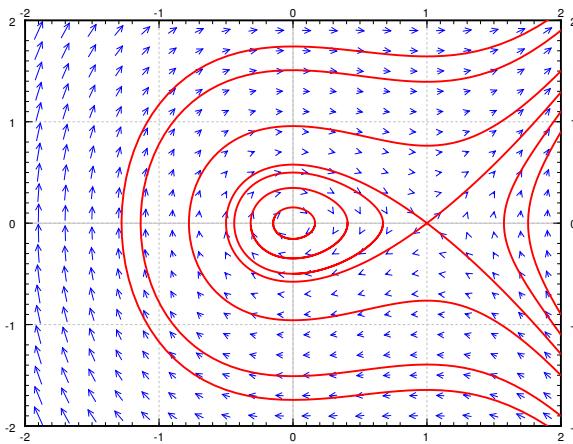


Figure 5.1: Phase portrait with some trajectories of $x' = y$, $y' = -x + x^2$.

Let us concentrate on those points in the phase diagram above where the trajectories seem to start, end, or go around. We see two such points: $(0, 0)$ and $(1, 0)$. The trajectories seem to go around the point $(0, 0)$, and they seem to either go in or out of the point $(1, 0)$. These points are precisely those points where the derivatives of both x and y are zero. Let us define the *critical points* as the points (x, y) such that

$$\begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix} = \vec{0}.$$

In other words, these are the points where both $f(x, y) = 0$ and $g(x, y) = 0$.

The critical points are where the behavior of the system is in some sense the most complicated. If $\begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix}$ is zero, then nearby, the vector can point in any direction whatsoever. Also, the trajectories are either going towards, away from, or around these points, so if we are looking for long-term qualitative behavior of the system, we should look at what is happening near the critical points.

Critical points are also sometimes called *equilibria*, since we have so-called *equilibrium solutions* at critical points. If (x_0, y_0) is a critical point, then we have the solutions

$$x(t) = x_0, \quad y(t) = y_0.$$

In Example 5.1.1 on the facing page, there are two equilibrium solutions:

$$x(t) = 0, \quad y(t) = 0, \quad \text{and} \quad x(t) = 1, \quad y(t) = 0.$$

Compare this discussion on equilibria to the discussion in § 1.7. The underlying concept is exactly the same.

5.1.2 Linearization

For a linear system of two variables given by an invertible matrix, the only critical point is the origin $(0, 0)$. In § 4.5 we studied the behavior of a homogeneous linear system of two

equations near a critical point. Let us put the understanding we gained in that section to good use understanding what happens near critical points of nonlinear systems.

In calculus we learned to estimate a function by taking its derivative and linearizing. We work similarly with nonlinear systems of ODE. Suppose (x_0, y_0) is a critical point. In order to linearize the system of differential equations, we want to linearize the two functions $f(x, y)$ and $g(x, y)$ that define this system. To do so, we will replace f and g by the tangent plane approximation to the functions. That is, if we set $z = f(x, y)$, the tangent plane is given by

$$L_f(x, y) = f(x_0, y_0) + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0).$$

Since (x_0, y_0) is a critical point, we know that $f(x_0, y_0) = 0$, so the tangent plane is given by

$$L_f(x, y) = f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0).$$

Similarly, the tangent plane for $g(x, y)$ near the critical point (x_0, y_0) is given by

$$L_g(x, y) = g_x(x_0, y_0)(x - x_0) + g_y(x_0, y_0)(y - y_0).$$

The idea of linearization in calculus was that we could use the tangent line or tangent plane to approximate a function near to a given point. For systems of differential equations, the idea is that we can approximate the solutions to the system of differential equations by the solutions to the linearized systems as long as we stay near the critical point. That means that we can approximate the solution to

$$\begin{aligned} \frac{dx}{dt} &= f(x, y) \\ \frac{dy}{dt} &= g(x, y) \end{aligned}$$

near the critical point (x_0, y_0) by the solution to the system

$$\begin{aligned} \frac{dx}{dt} &= f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0) \\ \frac{dy}{dt} &= g_x(x_0, y_0)(x - x_0) + g_y(x_0, y_0)(y - y_0) \end{aligned}$$

Next, change variables to (u, v) , so that $(u, v) = (0, 0)$ corresponds to (x_0, y_0) . That is,

$$u = x - x_0, \quad v = y - y_0.$$

Since $\frac{dx}{dt} = \frac{du}{dt}$ and $\frac{dy}{dt} = \frac{dv}{dt}$, we can rewrite the approximation system as

$$\begin{aligned} \frac{du}{dt} &= f_x(x_0, y_0)u + f_y(x_0, y_0)v \\ \frac{dv}{dt} &= g_x(x_0, y_0)u + g_y(x_0, y_0)v \end{aligned}$$

In multivariable calculus you may have seen that the several variables version of the derivative is the *Jacobian matrix*^{*}. The Jacobian matrix of the vector-valued function $\begin{bmatrix} f(x,y) \\ g(x,y) \end{bmatrix}$ at (x_0, y_0) is

$$\begin{bmatrix} \frac{\partial f}{\partial x}(x_0, y_0) & \frac{\partial f}{\partial y}(x_0, y_0) \\ \frac{\partial g}{\partial x}(x_0, y_0) & \frac{\partial g}{\partial y}(x_0, y_0) \end{bmatrix}.$$

This matrix gives the best linear approximation as u and v (and therefore x and y) vary. We define the *linearization* of the equation (5.1) as the linear system

$$\begin{bmatrix} u \\ v \end{bmatrix}' = \begin{bmatrix} \frac{\partial f}{\partial x}(x_0, y_0) & \frac{\partial f}{\partial y}(x_0, y_0) \\ \frac{\partial g}{\partial x}(x_0, y_0) & \frac{\partial g}{\partial y}(x_0, y_0) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}.$$

Example 5.1.2: Analyze the critical points for the system of differential equations in Example 5.1.1: $x' = y$, $y' = -x + x^2$.

Solution: There are two critical points, $(0, 0)$ and $(1, 0)$. The Jacobian matrix at any point is

$$\begin{bmatrix} \frac{\partial f}{\partial x}(x, y) & \frac{\partial f}{\partial y}(x, y) \\ \frac{\partial g}{\partial x}(x, y) & \frac{\partial g}{\partial y}(x, y) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 + 2x & 0 \end{bmatrix}.$$

Therefore at $(0, 0)$, we have $u = x$ and $v = y$, and the linearization is

$$\begin{bmatrix} u \\ v \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}.$$

At the point $(1, 0)$, we have $u = x - 1$ and $v = y$, and the linearization is

$$\begin{bmatrix} u \\ v \end{bmatrix}' = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}.$$

The phase diagrams of the two linearizations at the point $(0, 0)$ and $(1, 0)$ are given in Figure 5.2 on the next page. Note that the variables are now u and v . Compare Figure 5.2 with Figure 5.1 on page 315, and look especially at the behavior near the critical points. \square

5.1.3 Exercises

Exercise 5.1.1: Sketch the phase plane vector field for:

a) $x' = x^2$, $y' = y^2$, b) $x' = (x - y)^2$, $y' = -x$, c) $x' = e^y$, $y' = e^x$.

Exercise 5.1.2: Find the critical points and linearizations of the following systems.

a) $x' = x^2 - y^2$, $y' = x^2 + y^2 - 1$, b) $x' = -y$, $y' = 3x + yx^2$,
 c) $x' = x^2 + y$, $y' = y^2 + x$.

*Named for the German mathematician Carl Gustav Jacob Jacobi (1804–1851).

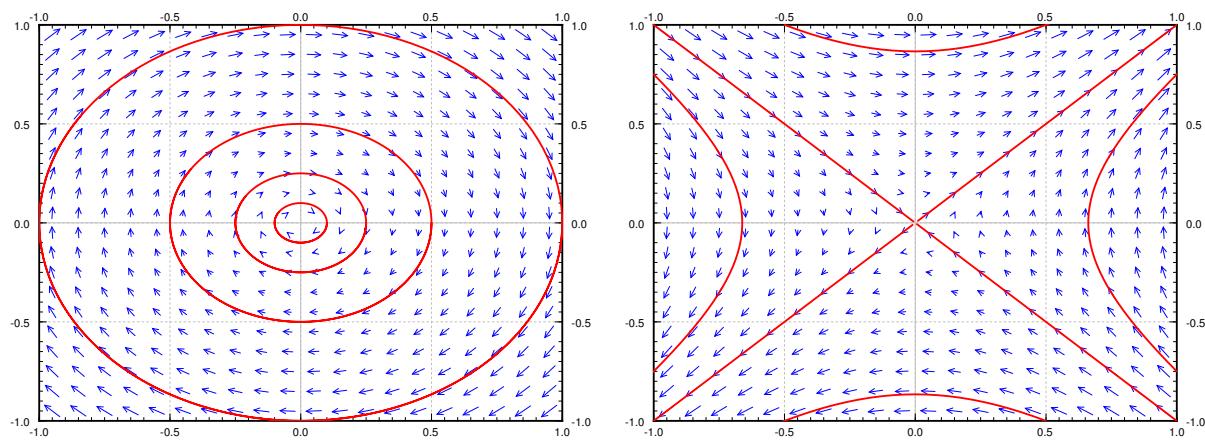


Figure 5.2: Phase diagram with some trajectories of linearizations at the critical points $(0,0)$ (left) and $(1,0)$ (right) of $x' = y$, $y' = -x + x^2$.

Exercise 5.1.3:* Find the critical points and linearizations of the following systems.

- a) $x' = \sin(\pi y) + (x-1)^2$, $y' = y^2 - y$,
- b) $x' = x + y + y^2$, $y' = x$,
- c) $x' = (x-1)^2 + y$, $y' = x^2 + y$.

Exercise 5.1.4: For the following systems, verify they have critical point at $(0,0)$, and find the linearization at $(0,0)$.

- a) $x' = x + 2y + x^2 - y^2$, $y' = 2y - x^2$
- b) $x' = -y$, $y' = x - y^3$
- c) $x' = ax + by + f(x,y)$, $y' = cx + dy + g(x,y)$, where $f(0,0) = 0$, $g(0,0) = 0$, and all first partial derivatives of f and g are also zero at $(0,0)$, that is, $\frac{\partial f}{\partial x}(0,0) = \frac{\partial f}{\partial y}(0,0) = \frac{\partial g}{\partial x}(0,0) = \frac{\partial g}{\partial y}(0,0) = 0$.

Exercise 5.1.5: Take the system $x' = (x-2)(x+y)$, $y' = (y+3)(x-y)$.

- a) Find all critical points.
- b) Determine the linearization of this system around each of the critical points.
- c) For each of the critical points, determine the behavior and classify the type of solution that the linearized system will have around that critical point.

Exercise 5.1.6: Take the system $x' = (x^2 - y)(x+3)$, $y' = (y-1)(x+y+1)$.

- a) Find all critical points.
- b) Determine the linearization of this system around each of the critical points.
- c) For each of the critical points, determine the behavior and classify the type of solution that the linearized system will have around that critical point.

Exercise 5.1.7: Take $x' = (x - y)^2$, $y' = (x + y)^2$.

- a) Find the set of critical points.
- b) Sketch a phase diagram and describe the behavior near the critical point(s).
- c) Find the linearization. Is it helpful in understanding the system?

Exercise 5.1.8: Take $x' = x^2$, $y' = x^3$.

- a) Find the set of critical points.
- b) Sketch a phase diagram and describe the behavior near the critical point(s).
- c) Find the linearization. Is it helpful in understanding the system?

Exercise 5.1.9:* The idea of critical points and linearization works in higher dimensions as well. You simply make the Jacobian matrix bigger by adding more functions and more variables. For the following system of 3 equations find the critical points and their linearizations:

$$x' = x + z^2, \quad y' = z^2 - y, \quad z' = z + x^2.$$

Exercise 5.1.10:* Any two-dimensional non-autonomous system $x' = f(x, y, t)$, $y' = g(x, y, t)$ can be written as a three-dimensional autonomous system (three equations). Write down this autonomous system using the variables u , v , w .

5.2 Stability and classification of isolated critical points

Attribution: [JL], §8.2.

Learning Objectives

After this section, you will be able to:

- Determine if a critical point of a non-linear system is isolated,
- Use the Jacobian matrix to classify the critical point of a non-linear system,
- Determine the stability of a critical point from the classification,
- Find the trajectories for a non-linear system, and
- Determine if a system is Hamiltonian and use that fact to find the general solution.

5.2.1 Isolated critical points and almost linear systems

A critical point is *isolated* if it is the only critical point in some small “neighborhood” of the point. That is, if we zoom in far enough it is the only critical point we see. In the example above, the critical point was isolated. If on the other hand there would be a whole curve of critical points, then it would not be isolated.

A system is called *almost linear* at a critical point (x_0, y_0) , if the critical point is isolated and the Jacobian matrix at the point is invertible, or equivalently if the linearized system has an isolated critical point. This is also equivalent to zero not being an eigenvalue of the Jacobian matrix at the critical point. In such a case, the nonlinear terms are very small and the system behaves like its linearization, at least if we are close to the critical point.

For example, the system in Examples 5.1.1 and 5.1.2 has two isolated critical points $(0, 0)$ and $(0, 1)$, and is almost linear at both critical points as the Jacobian matrices at both points, $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, are invertible.

On the other hand, the system $x' = x^2$, $y' = y^2$ has an isolated critical point at $(0, 0)$, however the Jacobian matrix

$$\begin{bmatrix} 2x & 0 \\ 0 & 2y \end{bmatrix}$$

is zero when $(x, y) = (0, 0)$. So the system is not almost linear. Even a worse example is the system $x' = x$, $y' = x^2$, which does not have isolated critical points; x' and y' are both zero whenever $x = 0$, that is, the entire y -axis.

Fortunately, most often critical points are isolated, and the system is almost linear at the critical points. So if we learn what happens there, we will have figured out the majority of situations that arise in applications.

5.2.2 Stability and classification of isolated critical points

Once we have an isolated critical point, the system is almost linear at that critical point, and we computed the associated linearized system, we can classify what happens to the solutions. We more or less use the classification for linear two-variable systems from § 4.5, with one

minor caveat. Let us list the behaviors depending on the eigenvalues of the Jacobian matrix at the critical point in [Table 5.1](#). This table is very similar to [Table 4.1](#) on page 279, with the exception of missing “center” points. The repeated eigenvalue cases are also missing. They behave similarly to the real eigenvalue descriptions in the table below, but similar to centers, the behavior can change slightly. It can behave like either a spiral or a node, but will be either a source or sink based on the sign of the repeated eigenvalue. We will discuss centers later, as they are more complicated.

Eigenvalues of the Jacobian matrix	Behavior	Stability
real and both positive	source / unstable node	unstable
real and both negative	sink / stable node	asymptotically stable
real and opposite signs	saddle	unstable
complex with positive real part	spiral source	unstable
complex with negative real part	spiral sink	asymptotically stable

Table 5.1: Behavior of an almost linear system near an isolated critical point.

In the third column, we mark points as *asymptotically stable* or *unstable*. Formally, a *stable critical point* (x_0, y_0) is one where given any small distance ϵ to (x_0, y_0) , and any initial condition within a perhaps smaller radius around (x_0, y_0) , the trajectory of the system never goes further away from (x_0, y_0) than ϵ . An *unstable critical point* is one that is not stable. Informally, a point is stable if we start close to a critical point and follow a trajectory we either go towards, or at least not away from, this critical point.

A stable critical point (x_0, y_0) is called *asymptotically stable* if given any initial condition sufficiently close to (x_0, y_0) and any solution $(x(t), y(t))$ satisfying that condition, then

$$\lim_{t \rightarrow \infty} (x(t), y(t)) = (x_0, y_0).$$

That is, the critical point is asymptotically stable if any trajectory for a sufficiently close initial condition goes towards the critical point (x_0, y_0) .

Example 5.2.1: Find and analyze the critical points of $x' = -y - x^2$, $y' = -x + y^2$.

Solution: See [Figure 5.3](#) on the next page for the phase diagram. Let us find the critical points. These are the points where $-y - x^2 = 0$ and $-x + y^2 = 0$. The first equation means $y = -x^2$, and so $y^2 = x^4$. Plugging into the second equation we obtain $-x + x^4 = 0$. Factoring we obtain $x(1 - x^3) = 0$. Since we are looking only for real solutions we get either $x = 0$ or $x = 1$. Solving for the corresponding y using $y = -x^2$, we get two critical points, one being $(0, 0)$ and the other being $(1, -1)$. Clearly the critical points are isolated.

Let us compute the Jacobian matrix:

$$\begin{bmatrix} -2x & -1 \\ -1 & 2y \end{bmatrix}.$$

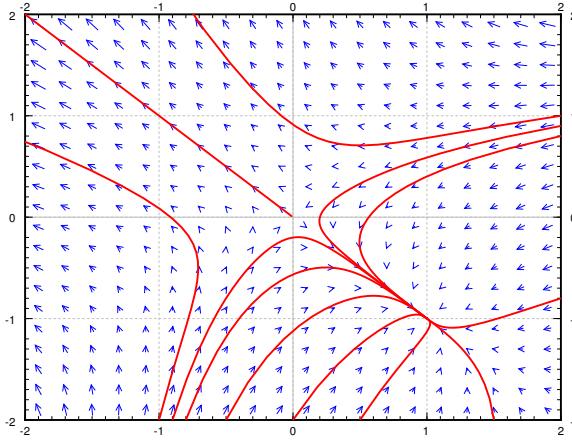


Figure 5.3: The phase portrait with few sample trajectories of $x' = -y - x^2$, $y' = -x + y^2$.

At the point $(0, 0)$ we get the matrix $\begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$ and so the two eigenvalues are 1 and -1 . As the matrix is invertible, the system is almost linear at $(0, 0)$. As the eigenvalues are real and of opposite signs, we get a saddle point, which is an unstable equilibrium point.

At the point $(1, -1)$ we get the matrix $\begin{bmatrix} -2 & -1 \\ -1 & -2 \end{bmatrix}$ and computing the eigenvalues we get -1 , -3 . The matrix is invertible, and so the system is almost linear at $(1, -1)$. As we have real eigenvalues and both negative, the critical point is a sink, and therefore an asymptotically stable equilibrium point. That is, if we start with any point (x_i, y_i) close to $(1, -1)$ as an initial condition and plot a trajectory, it approaches $(1, -1)$. In other words,

$$\lim_{t \rightarrow \infty} (x(t), y(t)) = (1, -1).$$

As you can see from the diagram, this behavior is true even for some initial points quite far from $(1, -1)$, but it is definitely not true for all initial points. \square

Example 5.2.2: Find and analyze the critical points of $x' = y + y^2 e^x$, $y' = x$.

Solution: First let us find the critical points. These are the points where $y + y^2 e^x = 0$ and $x = 0$. Simplifying we get $0 = y + y^2 = y(y + 1)$. So the critical points are $(0, 0)$ and $(0, -1)$, and hence are isolated. Let us compute the Jacobian matrix:

$$\begin{bmatrix} y^2 e^x & 1 + 2ye^x \\ 1 & 0 \end{bmatrix}.$$

At the point $(0, 0)$ we get the matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and so the two eigenvalues are 1 and -1 . As the matrix is invertible, the system is almost linear at $(0, 0)$. And, as the eigenvalues are real and of opposite signs, we get a saddle point, which is an unstable equilibrium point.

At the point $(0, -1)$ we get the matrix $\begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}$ whose eigenvalues are $\frac{1}{2} \pm i\frac{\sqrt{3}}{2}$. The matrix is invertible, and so the system is almost linear at $(0, -1)$. As we have complex eigenvalues with positive real part, the critical point is a spiral source, and therefore an unstable equilibrium point.

See Figure 5.4 on the facing page for the phase diagram. Notice the two critical points, and the behavior of the arrows in the vector field around these points. \square

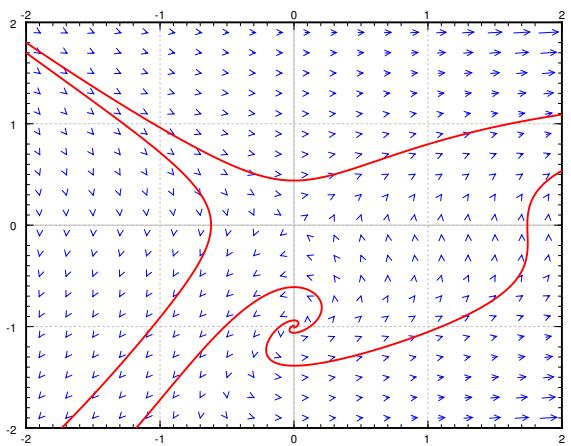


Figure 5.4: The phase portrait with few sample trajectories of $x' = y + y^2 e^x$, $y' = x$.

5.2.3 The trouble with centers

Recall, a linear system with a center means that trajectories travel in closed elliptical orbits in some direction around the critical point. Such a critical point we call a *center* or a *stable center*. It is not an asymptotically stable critical point, as the trajectories never approach the critical point, but at least if you start sufficiently close to the critical point, you stay close to the critical point. The simplest example of such behavior is the linear system with a center. Another example is the critical point $(0, 0)$ in Example 5.1.1 on page 314.

The trouble with a center in a nonlinear system is that whether the trajectory goes towards or away from the critical point is governed by the sign of the real part of the eigenvalues of the Jacobian matrix, and the Jacobian matrix in a nonlinear system changes from point to point. Since this real part is zero at the critical point itself, it can have either sign nearby, meaning the trajectory could be pulled towards or away from the critical point.

Example 5.2.3: Find and analyze the critical point(s) of $x' = y$, $y' = -x + y^3$.

Solution: The only critical point is the origin $(0, 0)$. The Jacobian matrix is

$$\begin{bmatrix} 0 & 1 \\ -1 & 3y^2 \end{bmatrix}.$$

At $(0, 0)$ the Jacobian matrix is $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, which has eigenvalues $\pm i$. So the linearization has a center.

Using the quadratic equation, the eigenvalues of the Jacobian matrix at any point (x, y) are

$$\lambda = \frac{3}{2}y^2 \pm i\frac{\sqrt{4 - 9y^4}}{2}.$$

At any point where $y \neq 0$ (so at most points near the origin), the eigenvalues have a positive real part (y^2 can never be negative). This positive real part pulls the trajectory away from the origin. A sample trajectory for an initial condition near the origin is given in Figure 5.5 on the following page.

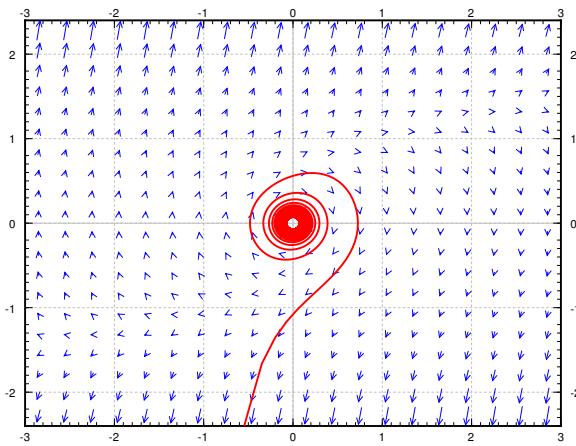


Figure 5.5: An unstable critical point (spiral source) at the origin for $x' = y, y' = -x + y^3$, even if the linearization has a center.

The same process could be carried out with the system $x' = y, y' = -x - y^3$. This one will also have a center as the linearization at the origin, but the non-linear system will have a spiral sink at the origin. The moral of the example is that further analysis is needed when the linearization has a center. The analysis will in general be more complicated than in the example above, and is more likely to involve case-by-case consideration. Such a complication should not be surprising to you. By now in your mathematical career, you have seen many places where a simple test is inconclusive, recall for example the second derivative test for maxima or minima, and requires more careful, and perhaps ad hoc analysis of the situation.

5.2.4 Conservative equations

An equation of the form

$$x'' + f(x) = 0$$

for an arbitrary function $f(x)$ is called a *conservative equation*. For example the pendulum equation is a conservative equation. The equations are conservative as there is no friction in the system so the energy in the system is “conserved.” Let us write this equation as a system of nonlinear ODE.

$$x' = y, \quad y' = -f(x).$$

These types of equations have the advantage that we can solve for their trajectories easily.

The trick is to first think of y as a function of x for a moment. Then use the chain rule

$$x'' = y' = \frac{dy}{dx}x' = y\frac{dy}{dx},$$

where the prime indicates a derivative with respect to t . We obtain $y\frac{dy}{dx} + f(x) = 0$. We

integrate with respect to x to get $\int y \frac{dy}{dx} dx + \int f(x) dx = C$. In other words

$$\frac{1}{2}y^2 + \int f(x) dx = C.$$

We obtained an implicit equation for the trajectories, with different C giving different trajectories. The value of C is conserved on any trajectory. This expression is sometimes called the *Hamiltonian* or the energy of the system. If you look back to § 1.8, you will notice that $y \frac{dy}{dx} + f(x) = 0$ is an exact equation, and we just found a potential function.

Example 5.2.4: Find the trajectories for the equation $x'' + x - x^2 = 0$, which is the equation from Example 5.1.1 on page 314.

Solution: The corresponding first order system is

$$x' = y, \quad y' = -x + x^2.$$

Trajectories satisfy

$$\frac{1}{2}y^2 + \frac{1}{2}x^2 - \frac{1}{3}x^3 = C.$$

We solve for y

$$y = \pm \sqrt{-x^2 + \frac{2}{3}x^3 + 2C}.$$

Plotting these graphs we get exactly the trajectories in Figure 5.1 on page 315. In particular we notice that near the origin the trajectories are *closed curves*: they keep going around the origin, never spiraling in or out. Therefore we discovered a way to verify that the critical point at $(0, 0)$ is a stable center. The critical point at $(0, 1)$ is a saddle as we already noticed. This example is typical for conservative equations. \square

Consider an arbitrary conservative equation $x'' + f(x) = 0$. All critical points occur when $y = 0$ (the x -axis), that is when $x' = 0$. The critical points are those points on the x -axis where $f(x) = 0$. The trajectories are given by

$$y = \pm \sqrt{-2 \int f(x) dx + 2C}.$$

So all trajectories are mirrored across the x -axis. In particular, there can be no spiral sources nor sinks. The Jacobian matrix is

$$\begin{bmatrix} 0 & 1 \\ -f'(x) & 0 \end{bmatrix}.$$

The critical point is almost linear if $f'(x) \neq 0$ at the critical point. Let J denote the Jacobian matrix. The eigenvalues of J are solutions to

$$0 = \det(J - \lambda I) = \lambda^2 + f'(x).$$

Therefore $\lambda = \pm \sqrt{-f'(x)}$. In other words, either we get real eigenvalues of opposite signs (if $f'(x) < 0$), or we get purely imaginary eigenvalues (if $f'(x) > 0$). There are only two possibilities for critical points, either an *unstable saddle point*, or a *stable center*. There are never any sinks or sources.

5.2.5 Hamiltonian Systems

A generalization of conservative equations to systems is a Hamiltonian system. This type of system has all of the nice properties of conservative equations when converted into systems, but allows for more general interactions between x and y . For these systems, the point is that the solution has a conserved quantity called a Hamiltonian, which does not change as the system evolves in time, which generally represents the energy of the system. Calling this function $H(x, y)$, this means that

$$\frac{d}{dt}H(x, y) = 0.$$

By the chain rule, this is equivalent to

$$\frac{\partial H}{\partial x}\frac{dx}{dt} + \frac{\partial H}{\partial y}\frac{dy}{dt} = 0.$$

One way to satisfy this is with

$$\begin{aligned}\frac{dx}{dt} &= -\frac{\partial H}{\partial y} \\ \frac{dy}{dt} &= \frac{\partial H}{\partial x}\end{aligned}\tag{5.2}$$

and this gives the definition of a *Hamiltonian system*.

Definition 5.2.1

The system

$$\begin{aligned}\frac{dx}{dt} &= f(x, y) \\ \frac{dy}{dt} &= g(x, y)\end{aligned}\tag{5.3}$$

is Hamiltonian if there is a function $H(x, y)$ so that $f(x, y) = -\frac{\partial H}{\partial y}$ and $g(x, y) = \frac{\partial H}{\partial x}$.

For solving these sorts of systems, we know that

$$\frac{d}{dt}H(x, y) = 0,$$

since that's how we defined the system. This means that the trajectories of this system are given by

$$H(x, y) = C$$

for a constant C determined by initial conditions. So if we can find the function H that expresses the system in the form (5.2), then we are done.

Finding this H is a lot similar to finding solutions to exact equations in § 1.8. First, we need to determine if the system is Hamiltonian. Since we want to have that

$$f(x, y) = -\frac{\partial H}{\partial y} \quad g(x, y) = \frac{\partial H}{\partial x}$$

we know that

$$f_x(x, y) = -\frac{\partial^2 H}{\partial x \partial y} \quad g_y(x, y) = \frac{\partial^2 H}{\partial x \partial y}$$

which shows that

$$f_x + g_y = 0.$$

This is what we can use to check if a system is Hamiltonian; compare to Theorem 1.8.1 for exact equations.

Once we know that a system is Hamiltonian, we can integrate the different components of the equation to find the function H . Since $f = -\frac{\partial H}{\partial y}$, then we can write

$$H(x, y) = - \int f(x, y) dy + A(x)$$

where $A(x)$ is an unknown function, which can be determined by differentiating this in x and setting equal to $g(x, y)$.

Example 5.2.5: Consider the system of differential equations given by

$$x' = -4x + 3y \quad y' = 2x + 4y.$$

Determine if this system is Hamiltonian and, if so, find the trajectories of the solution.

Solution: We first check if $f_x + g_y = 0$ to see if the system is Hamiltonian. Since $f_x = -4$ and $g_y = 4$, this means we have a Hamiltonian system. In order to find the function H , we use that

$$\frac{\partial H}{\partial y} = -f(x, y) = 4x - 3y.$$

Integrating both sides in y gives that

$$H(x, y) = 4xy - \frac{3}{2}y^2 + A(x)$$

for an unknown function $A(x)$. Differentiating this in x gives

$$\frac{\partial H}{\partial x} = 4y + A'(x)$$

which we want to equal $2x + 4y$. This gives that $A'(x) = 2x$ so $A(x) = x^2$. Thus, the Hamiltonian is given by

$$H(x, y) = x^2 + 4xy - \frac{3}{2}y^2$$

so that the trajectories are defined by

$$x^2 + 4xy - \frac{3}{2}y^2 = C$$

for any constant C . These are sketched in Figure 5.6. □

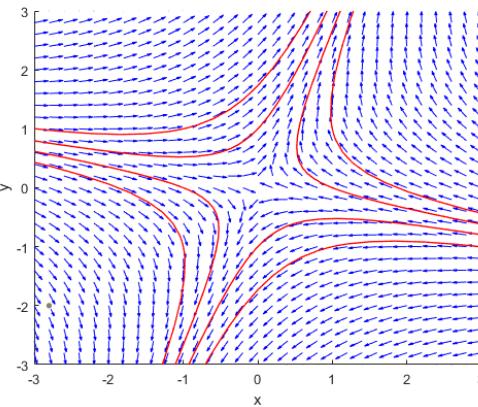


Figure 5.6: Vector field and trajectories for a Hamiltonian System

5.2.6 Exercises

Exercise 5.2.1: For the systems below, find and classify the critical points, also indicate if the equilibria are stable, asymptotically stable, or unstable.

a) $x' = -x + 3x^2, y' = -y$

b) $x' = x^2 + y^2 - 1, y' = x$

c) $x' = ye^x, y' = y - x + y^2$

Exercise 5.2.2:* For the systems below, find and classify the critical points.

a) $x' = -x + x^2, y' = y$

b) $x' = y - y^2 - x, y' = -x$

c) $x' = xy, y' = x + y - 1$

Exercise 5.2.3: Find and classify all critical points of the system

$$\frac{dx}{dt} = (x+1)(x-y+3) \quad \frac{dy}{dt} = (x-2)(x-y).$$

Exercise 5.2.4: Find and classify all critical points of the system

$$\frac{dx}{dt} = x^2 - y^2 \quad \frac{dy}{dt} = (x+4)(y-2).$$

Exercise 5.2.5: Find the implicit equations of the trajectories of the following conservative systems. Next find their critical points (if any) and classify them.

a) $x'' + x + x^3 = 0$

b) $\theta'' + \sin \theta = 0$

c) $z'' + (z-1)(z+1) = 0$

d) $x'' + x^2 + 1 = 0$

Exercise 5.2.6:* Find the implicit equations of the trajectories of the following conservative systems. Next find their critical points (if any) and classify them.

a) $x'' + x^2 = 4$

b) $x'' + e^x = 0$

c) $x'' + (x+1)e^x = 0$

Exercise 5.2.7: Find and classify the critical point(s) of $x' = -x^2$, $y' = -y^2$.

Exercise 5.2.8: Suppose $x' = -xy$, $y' = x^2 - 1 - y$.

- a) Show there are two spiral sinks at $(-1, 0)$ and $(1, 0)$.
- b) For any initial point of the form $(0, y_0)$, find what is the trajectory.
- c) Can a trajectory starting at (x_0, y_0) where $x_0 > 0$ spiral into the critical point at $(-1, 0)$? Why or why not?

Exercise 5.2.9:* The conservative system $x'' + x^3 = 0$ is not almost linear. Classify its critical point(s) nonetheless.

Exercise 5.2.10: Determine if the following system is Hamiltonian. If it is, find the general solution in the form $H(x, y) = C$ and sketch some of the trajectories.

$$\frac{dx}{dt} = x - 2y \quad \frac{dy}{dt} = 3x - y.$$

Exercise 5.2.11: Determine if the following system is Hamiltonian. If it is, find the general solution in the form $H(x, y) = C$ and sketch some of the trajectories.

$$\frac{dx}{dt} = 4x - 2y + 2 \quad \frac{dy}{dt} = -5x + y - 1.$$

Exercise 5.2.12: Determine if the following system is Hamiltonian. If it is, find the general solution in the form $H(x, y) = C$ and sketch some of the trajectories.

$$\frac{dx}{dt} = x^2 - 2xy + 3y^2 \quad \frac{dy}{dt} = y^2 - 2xy + e^x.$$

Exercise 5.2.13: Determine if the following system is Hamiltonian. If it is, find the general solution in the form $H(x, y) = C$ and sketch some of the trajectories.

$$\frac{dx}{dt} = 3x - 2xy \quad \frac{dy}{dt} = 2xy - 3y.$$

Exercise 5.2.14: In the example $x' = y$, $y' = y^3 - x$ show that for any trajectory, the distance from the origin is an increasing function. Conclude that the origin behaves like a spiral source. Hint: Consider $f(t) = (x(t))^2 + (y(t))^2$ and show it has positive derivative.

Exercise 5.2.15:* Derive an analogous classification of critical points for equations in one dimension, such as $x' = f(x)$ based on the derivative. A point x_0 is critical when $f(x_0) = 0$ and almost linear if in addition $f'(x_0) \neq 0$. Figure out if the critical point is stable or unstable depending on the sign of $f'(x_0)$. Explain. Hint: see § 1.7.

Exercise 5.2.16: Suppose f is always positive. Find the trajectories of $x'' + f(x') = 0$. Are there any critical points?

Exercise 5.2.17: Suppose that $x' = f(x, y)$, $y' = g(x, y)$. Suppose that $g(x, y) > 1$ for all x and y . Are there any critical points? What can we say about the trajectories at t goes to infinity?

5.3 Applications of nonlinear systems

Attribution: [JL], §8.3.

Learning Objectives

After this section, you will be able to:

- Use non-linear systems to model the motion of a pendulum, and
- Use non-linear systems to model population dynamics like predator-prey and competing species models.

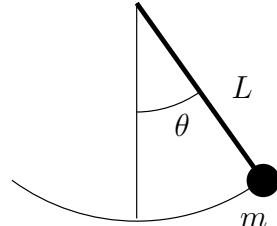
In this section we study two very standard examples of nonlinear systems. First, we look at the nonlinear pendulum equation. We saw the pendulum equation's linearization before, but we noted it was only valid for small angles and short times. Now we find out what happens for large angles. Next, we look at the predator-prey equation, which finds various applications in modeling problems in biology, chemistry, economics, and elsewhere.

5.3.1 Pendulum

The first example we study is the pendulum equation $\theta'' + \frac{g}{L} \sin \theta = 0$. Here, θ is the angular displacement, g is the gravitational acceleration, and L is the length of the pendulum. In this equation we disregard friction, so we are talking about an idealized pendulum.

This equation is a conservative equation, so we can use our analysis of conservative equations from the previous section. Let us change the equation to a two-dimensional system in variables (θ, ω) by introducing the new variable ω :

$$\begin{bmatrix} \theta \\ \omega \end{bmatrix}' = \begin{bmatrix} \omega \\ -\frac{g}{L} \sin \theta \end{bmatrix}.$$



The critical points of this system are when $\omega = 0$ and $-\frac{g}{L} \sin \theta = 0$, or in other words if $\sin \theta = 0$. So the critical points are when $\omega = 0$ and θ is a multiple of π . That is, the points are $\dots(-2\pi, 0), (-\pi, 0), (0, 0), (\pi, 0), (2\pi, 0) \dots$. While there are infinitely many critical points, they are all isolated. Let us compute the Jacobian matrix:

$$\begin{bmatrix} \frac{\partial}{\partial \theta}(\omega) & \frac{\partial}{\partial \omega}(\omega) \\ \frac{\partial}{\partial \theta}\left(-\frac{g}{L} \sin \theta\right) & \frac{\partial}{\partial \omega}\left(-\frac{g}{L} \sin \theta\right) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{g}{L} \cos \theta & 0 \end{bmatrix}.$$

For conservative equations, there are two types of critical points. Either stable centers, or saddle points. The eigenvalues of the Jacobian matrix are $\lambda = \pm \sqrt{-\frac{g}{L}} \cos \theta$.

The eigenvalues are going to be real when $\cos \theta < 0$. This happens at the odd multiples of π . The eigenvalues are going to be purely imaginary when $\cos \theta > 0$. This happens at the even multiples of π . Therefore the system has a stable center at the points $\dots(-2\pi, 0), (0, 0), (2\pi, 0) \dots$, and it has an unstable saddle at the points $\dots(-3\pi, 0), (-\pi, 0), (\pi, 0), (3\pi, 0) \dots$. Look at the phase diagram in [Figure 5.7](#) on the next page, where for simplicity we let $\frac{g}{L} = 1$.

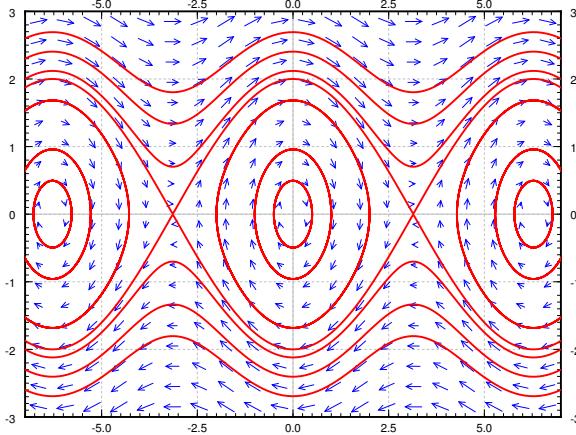


Figure 5.7: Phase plane diagram and some trajectories of the nonlinear pendulum equation.

In the linearized equation we have only a single critical point, the center at $(0, 0)$. Now we see more clearly what we meant when we said the linearization is good for small angles. The horizontal axis is the deflection angle. The vertical axis is the angular velocity of the pendulum. Suppose we start at $\theta = 0$ (no deflection), and we start with a small angular velocity ω . Then the trajectory keeps going around the critical point $(0, 0)$ in an approximate circle. This corresponds to short swings of the pendulum back and forth. When θ stays small, the trajectories really look like circles and hence are very close to our linearization.

When we give the pendulum a big enough push, it goes across the top and keeps spinning about its axis. This behavior corresponds to the wavy curves that do not cross the horizontal axis in the phase diagram. Let us suppose we look at the top curves, when the angular velocity ω is large and positive. Then the pendulum is going around and around its axis. The velocity is going to be large when the pendulum is near the bottom, and the velocity is the smallest when the pendulum is close to the top of its loop.

At each critical point, there is an equilibrium solution. Consider the solution $\theta = 0$; the pendulum is not moving and is hanging straight down. This is a stable place for the pendulum to be, hence this is a *stable* equilibrium.

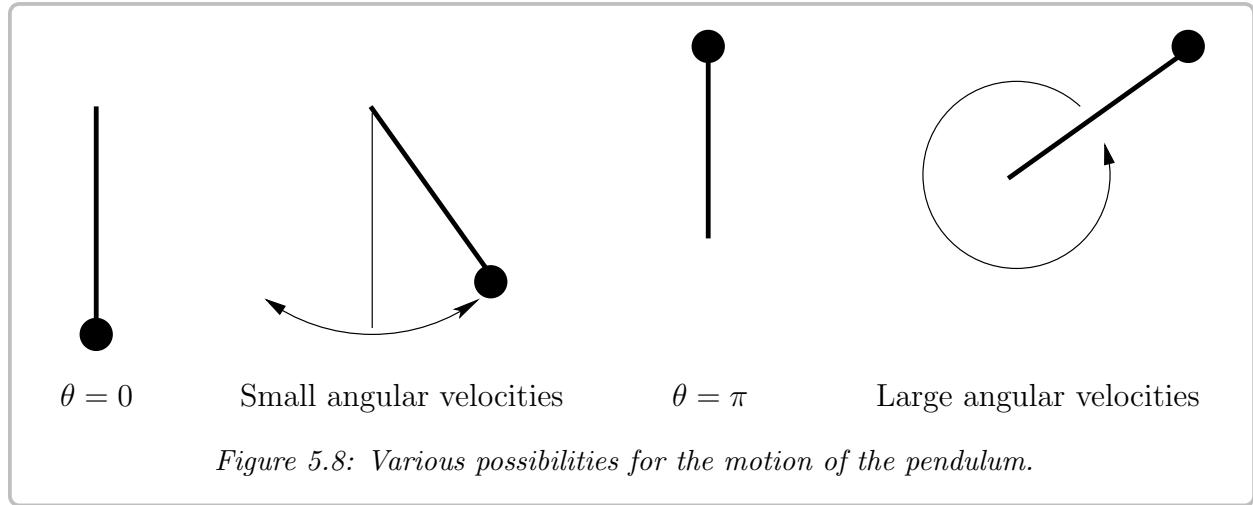
The other type of equilibrium solution is at the unstable point, for example $\theta = \pi$. Here the pendulum is upside down. Sure you can balance the pendulum this way and it will stay, but this is an *unstable* equilibrium. Even the tiniest push will make the pendulum start swinging wildly.

See Figure 5.8 on the following page for a diagram. The first picture is the stable equilibrium $\theta = 0$. The second picture corresponds to those “almost circles” in the phase diagram around $\theta = 0$ when the angular velocity is small. The next picture is the unstable equilibrium $\theta = \pi$. The last picture corresponds to the wavy lines for large angular velocities.

The quantity

$$\frac{1}{2}\omega^2 - \frac{g}{L} \cos \theta$$

is conserved by any solution. This is the energy or the Hamiltonian of the system.



We have a conservative equation and so (exercise) the trajectories are given by

$$\omega = \pm \sqrt{\frac{2g}{L} \cos \theta + C},$$

for various values of C . Let us look at the initial condition of $(\theta_0, 0)$, that is, we take the pendulum to angle θ_0 , and just let it go (initial angular velocity 0). We plug the initial conditions into the above and solve for C to obtain

$$C = -\frac{2g}{L} \cos \theta_0.$$

Thus the expression for the trajectory is

$$\omega = \pm \sqrt{\frac{2g}{L} \sqrt{\cos \theta - \cos \theta_0}}.$$

Let us figure out the period. That is, the time it takes for the pendulum to swing back and forth. We notice that the trajectory about the origin in the phase plane is symmetric about both the θ and the ω -axis. That is, in terms of θ , the time it takes from θ_0 to $-\theta_0$ is the same as it takes from $-\theta_0$ back to θ_0 . Furthermore, the time it takes from $-\theta_0$ to 0 is the same as to go from 0 to θ_0 . Therefore, let us find how long it takes for the pendulum to go from angle 0 to angle θ_0 , which is a quarter of the full oscillation and then multiply by 4.

We figure out this time by finding $\frac{dt}{d\theta}$ and integrating from 0 to θ_0 . The period is four times this integral. Let us stay in the region where ω is positive. Since $\omega = \frac{d\theta}{dt}$, inverting we get

$$\frac{dt}{d\theta} = \sqrt{\frac{L}{2g} \frac{1}{\sqrt{\cos \theta - \cos \theta_0}}}.$$

Therefore the period T is given by

$$T = 4 \sqrt{\frac{L}{2g}} \int_0^{\theta_0} \frac{1}{\sqrt{\cos \theta - \cos \theta_0}} d\theta.$$

The integral is an improper integral, and we cannot in general evaluate it symbolically. We must resort to numerical approximation if we want to compute a particular T .

Recall from § 2.4, the linearized equation $\theta'' + \frac{g}{L}\theta = 0$ has period

$$T_{\text{linear}} = 2\pi\sqrt{\frac{L}{g}}.$$

We plot T , T_{linear} , and the relative error $\frac{T-T_{\text{linear}}}{T}$ in Figure 5.9. The relative error says how far is our approximation from the real period percentage-wise. Note that T_{linear} is simply a constant, it does not change with the initial angle θ_0 . The actual period T gets larger and larger as θ_0 gets larger. Notice how the relative error is small when θ_0 is small. It is still only 15% when $\theta_0 = \frac{\pi}{2}$, that is, a 90 degree angle. The error is 3.8% when starting at $\frac{\pi}{4}$, a 45 degree angle. At a 5 degree initial angle, the error is only 0.048%.

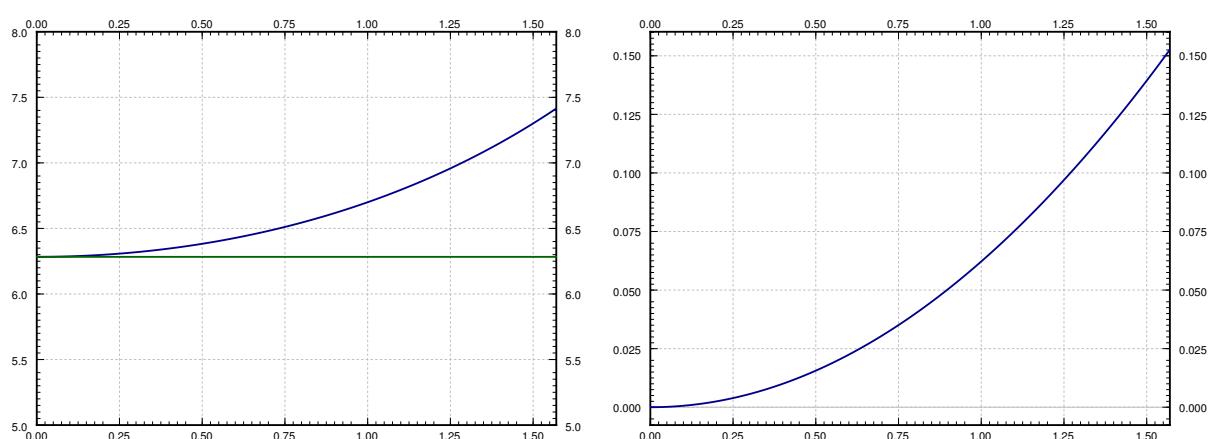


Figure 5.9: The plot of T and T_{linear} with $\frac{g}{L} = 1$ (left), and the plot of the relative error $\frac{T-T_{\text{linear}}}{T}$ (right), for θ_0 between 0 and $\pi/2$.

While it is not immediately obvious from the formula, it is true that

$$\lim_{\theta_0 \uparrow \pi} T = \infty.$$

That is, the period goes to infinity as the initial angle approaches the unstable equilibrium point. So if we put the pendulum almost upside down it may take a very long time before it gets down. This is consistent with the limiting behavior, where the exactly upside down pendulum never makes an oscillation, so we could think of that as infinite period.

5.3.2 Predator-prey or Lotka–Volterra systems

One of the most common simple applications of nonlinear systems are the so-called *predator-prey* or *Lotka–Volterra** systems. For example, these systems arise when two species interact,

*Named for the American mathematician, chemist, and statistician Alfred James Lotka (1880–1949) and the Italian mathematician and physicist Vito Volterra (1860–1940).

one as the prey and one as the predator. It is then no surprise that the equations also see applications in economics. The system also arises in chemical reactions. In biology, this system of equations explains the natural periodic variations of populations of different species in nature. Before the application of differential equations, these periodic variations in the population baffled biologists.

We keep with the classical example of hares and foxes in a forest, it is the easiest to understand.

$$\begin{aligned}x &= \# \text{ of hares (the prey)}, \\y &= \# \text{ of foxes (the predator)}.\end{aligned}$$

When there are a lot of hares, there is plenty of food for the foxes, so the fox population grows. However, when the fox population grows, the foxes eat more hares, so when there are lots of foxes, the hare population should go down, and vice versa. The Lotka–Volterra model proposes that this behavior is described by the system of equations

$$\begin{aligned}x' &= (a - by)x, \\y' &= (cx - d)y,\end{aligned}$$

where a, b, c, d are some parameters that describe the interaction of the foxes and hares*. In this model, these are all positive numbers.

Let us analyze the idea behind this model. The model is a slightly more complicated idea based on the exponential population model. First expand,

$$x' = (a - by)x = ax - byx.$$

The hares are expected to simply grow exponentially in the absence of foxes, that is where the ax term comes in, the growth in population is proportional to the population itself. We are assuming the hares always find enough food and have enough space to reproduce. However, there is another component $-byx$, that is, the population also is decreasing proportionally to the number of foxes. Together we can write the equation as $(a - by)x$, so it is like exponential growth or decay but the constant depends on the number of foxes.

The equation for foxes is very similar, expand again

$$y' = (cx - d)y = cxy - dy.$$

The foxes need food (hares) to reproduce: the more food, the bigger the rate of growth, hence the cxy term. On the other hand, there are natural deaths in the fox population, and hence the $-dy$ term.

Without further delay, let us start with an explicit example. Suppose the equations are

$$x' = (0.4 - 0.01y)x, \quad y' = (0.003x - 0.3)y.$$

See [Figure 5.10](#) on the facing page for the phase portrait. In this example it makes sense to also plot x and y as graphs with respect to time. Therefore the second graph in [Figure 5.10](#) is the graph of x and y on the vertical axis (the prey x is the thinner line with taller peaks), against time on the horizontal axis. The particular solution graphed was with initial conditions of 20 foxes and 50 hares.

*This interaction does not end well for the hare.

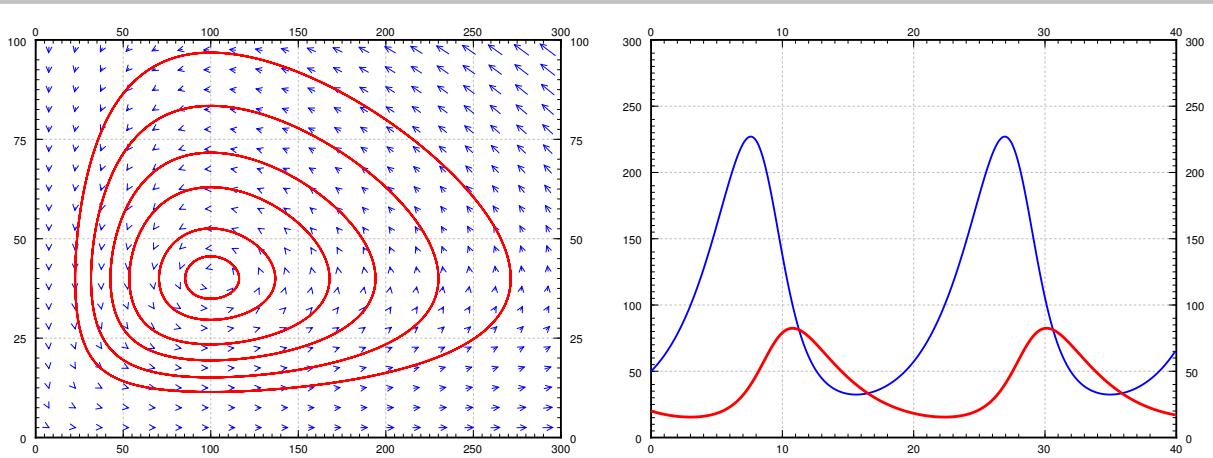


Figure 5.10: The phase portrait (left) and graphs of x and y for a sample solution (right).

Let us analyze what we see on the graphs. We work in the general setting rather than putting in specific numbers. We start with finding the critical points. Set $(a - by)x = 0$, and $(cx - d)y = 0$. The first equation is satisfied if either $x = 0$ or $y = a/b$. If $x = 0$, the second equation implies $y = 0$. If $y = a/b$, the second equation implies $x = d/c$. There are two equilibria: at $(0, 0)$ when there are no animals at all, and at $(d/c, a/b)$. In our specific example $x = d/c = 100$, and $y = a/b = 40$. This is the point where there are 100 hares and 40 foxes.

We compute the Jacobian matrix:

$$\begin{bmatrix} a - by & -bx \\ cy & cx - d \end{bmatrix}.$$

At the origin $(0, 0)$ we get the matrix $\begin{bmatrix} a & 0 \\ 0 & -d \end{bmatrix}$, so the eigenvalues are a and $-d$, hence real and of opposite signs. So the critical point at the origin is a saddle. This makes sense. If you started with some foxes but no hares, then the foxes would go extinct, that is, you would approach the origin. If you started with no foxes and a few hares, then the hares would keep multiplying without check, and so you would go away from the origin.

OK, how about the other critical point at $(d/c, a/b)$. Here the Jacobian matrix becomes

$$\begin{bmatrix} 0 & -\frac{bd}{c} \\ \frac{ac}{b} & 0 \end{bmatrix}.$$

The eigenvalues satisfy $\lambda^2 + ad = 0$. In other words, $\lambda = \pm i\sqrt{ad}$. The eigenvalues being purely imaginary, we are in the case where we cannot quite decide using only linearization. We could have a stable center, spiral sink, or a spiral source. That is, the equilibrium could be asymptotically stable, stable, or unstable. Of course I gave you a picture above that seems to imply it is a stable center. But never trust a picture only. Perhaps the oscillations are getting larger and larger, but only *very* slowly. Of course this would be bad as it would imply something will go wrong with our population sooner or later. And I only graphed a very specific example with very specific trajectories.

How can we be sure we are in the stable situation? As we said before, in the case of purely imaginary eigenvalues, we have to do a bit more work. Previously we found that for conservative systems, there was a certain quantity that was conserved on the trajectories, and hence the trajectories had to go in closed loops. We can use a similar technique here. We just have to figure out what is the conserved quantity. After some trial and error we find the constant

$$C = \frac{y^a x^d}{e^{cx+by}} = y^a x^d e^{-cx-by}$$

is conserved. Such a quantity is called the *constant of motion*. Let us check C really is a constant of motion. How do we check, you say? Well, a constant is something that does not change with time, so let us compute the derivative with respect to time:

$$C' = ay^{a-1}y'x^d e^{-cx-by} + y^a dx^{d-1}x'e^{-cx-by} + y^a x^d e^{-cx-by}(-cx' - by').$$

Our equations give us what x' and y' are so let us plug those in:

$$\begin{aligned} C' &= ay^{a-1}(cx-d)yx^d e^{-cx-by} + y^a dx^{d-1}(a-by)xe^{-cx-by} \\ &\quad + y^a x^d e^{-cx-by}(-c(a-by)x - b(cx-d)y) \\ &= y^a x^d e^{-cx-by} \left(a(cx-d) + d(a-by) + (-c(a-by)x - b(cx-d)y) \right) \\ &= 0. \end{aligned}$$

So along the trajectories C is constant. In fact, the expression $C = \frac{y^a x^d}{e^{cx+by}}$ gives us an implicit equation for the trajectories. In any case, once we have found this constant of motion, it must be true that the trajectories are simple curves, that is, the level curves of $\frac{y^a x^d}{e^{cx+by}}$. It turns out, the critical point at $(\frac{d}{c}, \frac{a}{b})$ is a maximum for C (left as an exercise). So $(\frac{d}{c}, \frac{a}{b})$ is a stable equilibrium point, and we do not have to worry about the foxes and hares going extinct or their populations exploding.

One blemish on this wonderful model is that the number of foxes and hares are discrete quantities and we are modeling with continuous variables. Our model has no problem with there being 0.1 fox in the forest for example, while in reality that makes no sense. The approximation is a reasonable one as long as the number of foxes and hares are large, but it does not make much sense for small numbers. One must be careful in interpreting any results from such a model.

An interesting consequence (perhaps counterintuitive) of this model is that adding animals to the forest might lead to extinction, because the variations will get too big, and one of the populations will get close to zero. For example, suppose there are 20 foxes and 50 hares as before, but now we bring in more foxes, bringing their number to 200. If we run the computation, we find the number of hares will plummet to just slightly more than 1 hare in the whole forest. In reality that most likely means the hares die out, and then the foxes will die out as well as they will have nothing to eat.

5.3.3 Competing Species systems

Another application of non-linear systems that also works with population models is a competing species interaction. The setup is that there are two species that live in the same

environment, and need to compete over resources. This means that both species will grow on their own, but when the two species interact, it is negative for both species. This gives rise to a system of differential equations of the form

$$\begin{aligned}\frac{dx}{dt} &= ax - bxy \\ \frac{dy}{dt} &= cy - dxy\end{aligned}$$

if both populations grow exponentially, or

$$\begin{aligned}\frac{dx}{dt} &= ax(K - x) - bxy \\ \frac{dy}{dt} &= cy(M - y) - dxy\end{aligned}$$

if both species grow logistically. The numbers here are all positive constants that explain how the different populations affect growth rates. For the logistic model, let's look at the equilibrium solutions. For this, we need

$$x(aK - ax - by) = 0 \quad y(cM - cy - dx) = 0$$

which gives equilibrium solutions at $(0, 0)$, $(0, M)$, and $(K, 0)$, all of which result in one (or both) of the species being extinct. The other equilibrium solution is more interesting, because it involves both species coexisting. This happens when

$$by = aK - ax \quad cy = cM - dx.$$

Solving this gives a critical point with $x > 0$ and $y > 0$.

The Jacobian matrix for this system is

$$J(x, y) = \begin{bmatrix} aK - 2ax - by & -bx \\ -dy & cM - 2cy - dx \end{bmatrix}.$$

Unlike the predator-prey system, there are multiple options for how this system can behave based on the values of a , b , c , d , K , and M . It is possible that the coexistence equilibrium solution will be a nodal sink, so that all nearby solutions will converge to it over time, and the species will continue to exist in harmony. However, it is also possible that the coexistence solution is a saddle and the solutions at $(K, 0)$ and $(0, M)$ are sinks. This means that coexistence is unstable, and that over time, the populations will converge to one of the other two equilibrium solutions, meaning that one of the species will die out as time goes on. Determining which will survive will require a numerical model since these equations can not be solved analytically.

Example 5.3.1: Analyze the competing species model given by the system of differential equations

$$x' = x(4 - x - 2y) \quad y' = y(7 - y - 3x).$$

Is the coexistence solution stable or unstable? What will happen to the populations over time?

Solution: Solving for the equilibrium solutions gives $(0, 0)$, $(4, 0)$, $(0, 7)$, and the coexistence solution where

$$4 - x = 2y \quad y = 7 - 3x.$$

Simplifying this gives

$$4 - x = 14 - 6x$$

or $x = 2$. The second equation then implies that $y = 1$.

The Jacobian for this system is

$$J(x, y) = \begin{bmatrix} 4 - 2x - 2y & -2x \\ -3y & 7 - 2y - 3x \end{bmatrix}.$$

Evaluating this matrix at the point $(2, 1)$ gives

$$\begin{bmatrix} -2 & -4 \\ -3 & -1 \end{bmatrix},$$

which we need to find the eigenvalues to classify what type of linearized solution we have here. These are determined by

$$(-2 - \lambda)(-1 - \lambda) - 12 = \lambda^2 + 3\lambda - 9 = 0.$$

Thus, the eigenvalues are given by

$$\lambda = \frac{-3 \pm \sqrt{9 + 36}}{2}$$

which will be real with opposite signs. Therefore, this equilibrium solution is a saddle, and unstable. To confirm this, we can also check the equilibrium solutions at $(4, 0)$ and $(0, 7)$. For $(4, 0)$, we get the matrix

$$\begin{bmatrix} -8 & -8 \\ 0 & -1 \end{bmatrix}$$

which is a nodal sink. For $(0, 7)$, we get

$$\begin{bmatrix} -10 & 0 \\ -21 & -7 \end{bmatrix}$$

which is also a nodal sink. Thus, we see that the coexistence equilibrium solution is unstable, and both of the equilibrium solutions with one species extinct are stable. Therefore, over time, one of the two species will die off depending on the initial population. \square

Showing that a system of equations has a stable solution can be a very difficult problem. When Isaac Newton put forth his laws of planetary motions, he proved that a single planet orbiting a single sun is a stable system. But any solar system with more than 1 planet proved very difficult indeed. In fact, such a system behaves chaotically (see § 5.5), meaning small changes in initial conditions lead to very different long-term outcomes. From numerical experimentation and measurements, we know the earth will not fly out into the empty space or crash into the sun, for at least some millions of years or so. But we do not know what happens beyond that.

5.3.4 Exercises

Exercise 5.3.1: Take the damped nonlinear pendulum equation $\theta'' + \mu\theta' + (g/L)\sin\theta = 0$ for some $\mu > 0$ (that is, there is some friction).

- Suppose $\mu = 1$ and $g/L = 1$ for simplicity, find and classify the critical points.
- Do the same for any $\mu > 0$ and any g and L , but such that the damping is small, in particular, $\mu^2 < 4(g/L)$.
- Explain what your findings mean, and if it agrees with what you expect in reality.

Exercise 5.3.2:* Take the damped nonlinear pendulum equation $\theta'' + \mu\theta' + (g/L)\sin\theta = 0$ for some $\mu > 0$ (that is, there is friction). Suppose the friction is large, in particular $\mu^2 > 4(g/L)$.

- Find and classify the critical points.
- Explain what your findings mean, and if it agrees with what you expect in reality.

Exercise 5.3.3: Suppose the hares do not grow exponentially, but logistically. In particular consider

$$x' = (0.4 - 0.01y)x - \gamma x^2, \quad y' = (0.003x - 0.3)y.$$

For the following two values of γ , find and classify all the critical points in the positive quadrant, that is, for $x \geq 0$ and $y \geq 0$. Then sketch the phase diagram. Discuss the implication for the long term behavior of the population.

- $\gamma = 0.001$,
- $\gamma = 0.01$.

Exercise 5.3.4:* Suppose we have the system predator-prey system where the foxes are also killed at a constant rate h (h foxes killed per unit time): $x' = (a - by)x$, $y' = (cx - d)y - h$.

- Find the critical points and the Jacobian matrices of the system.
- Put in the constants $a = 0.4$, $b = 0.01$, $c = 0.003$, $d = 0.3$, $h = 10$. Analyze the critical points. What do you think it says about the forest?

Exercise 5.3.5 (challenging):* Suppose the foxes never die. That is, we have the system $x' = (a - by)x$, $y' = cxy$. Find the critical points and notice they are not isolated. What will happen to the population in the forest if it starts at some positive numbers. Hint: Think of the constant of motion.

Exercise 5.3.6: The following system of differential equations models a pair of populations interacting.

$$\frac{dx}{dt} = 4x - 2xy \quad \frac{dy}{dt} = 3xy - y$$

- Does this system of differential equations better fit with a competing species model or a predator-prey model? If it is predator-prey, which species is the predator?
- Find and classify the critical point (if it exists) with both $x > 0$ and $y > 0$.
- Describe what is going to happen to the population of these species over time. It this depends on the initial condition, say so.

Exercise 5.3.7: The following system of differential equations models a pair of populations interacting.

$$\frac{dx}{dt} = x(6 - 3y - 2x) \quad \frac{dy}{dt} = y(4 - y - 3x)$$

- a) Does this system of differential equations better fit with a competing species model or a predator-prey model? If it is predator-prey, which species is the predator?
- b) Find and classify the critical point (if it exists) with both $x > 0$ and $y > 0$.
- c) Describe what is going to happen to the population of these species over time. It this depends on the initial condition, say so.

Exercise 5.3.8: The following system of differential equations models a pair of populations interacting.

$$\frac{dx}{dt} = x(5 - x - 2y) \quad \frac{dy}{dt} = y(7 - x - 3y)$$

- a) Does this system of differential equations better fit with a competing species model or a predator-prey model? If it is predator-prey, which species is the predator?
- b) Find and classify the critical point (if it exists) with both $x > 0$ and $y > 0$.
- c) Describe what is going to happen to the population of these species over time. It this depends on the initial condition, say so.

Exercise 5.3.9:

- a) Suppose x and y are positive variables. Show $\frac{yx}{e^{x+y}}$ attains a maximum at $(1, 1)$.
- b) Suppose a, b, c, d are positive constants, and also suppose x and y are positive variables. Show $\frac{y^a x^d}{e^{cx+by}}$ attains a maximum at $(d/c, a/b)$.

Exercise 5.3.10: Suppose that for the pendulum equation we take a trajectory giving the spinning-around motion, for example $\omega = \sqrt{\frac{2g}{L} \cos \theta + \frac{2g}{L} + \omega_0^2}$. This is the trajectory where the lowest angular velocity is ω_0^2 . Find an integral expression for how long it takes the pendulum to go all the way around.

Exercise 5.3.11: Consider a predator-prey interaction where humans have gotten involved. The idea is that at least one of the species is valuable for food or another resource, and the two species still intact in their normal predator-prey manner. The first version of this will deal with “constant effort harvesting,” which means that humans will remove animals from the populations at a rate proportional to the population. This results in equations of the form

$$\frac{dx}{dt} = x(a - by - E_1) \quad \frac{dy}{dt} = y(-d + cx - E_2)$$

where E_1 and E_2 denote the amount of harvesting done.

- a) There is a single equilibrium solution with $x > 0$ and $y > 0$ in the case of no harvesting, that is, $E_1 = E_2 = 0$. Find this equilibrium solution.
- b) Without doing any mathematical work, what do you think will happen to the equilibrium solution if just the prey is harvested? What if just the predator is harvested? What if both are harvested?
- c) Find the location of the equilibrium system in each of the three cases in the previous part. Do this in terms of the constants E_1 and E_2 for all three cases.

Exercise 5.3.12: The second version of this will deal with “constant yield harvesting,” which means that humans will remove animals from the populations at a fixed rate, no matter their population. This results in equations of the form

$$\frac{dx}{dt} = x(a - by) - H_1 \quad \frac{dy}{dt} = y(-d + cx) - H_2$$

where H_1 and H_2 denote the amount of harvesting done.

- a) There is a single equilibrium solution with $x > 0$ and $y > 0$ in the case of no harvesting, that is, $H_1 = H_2 = 0$. Find this equilibrium solution.
- b) Without doing any mathematical work, what do you think will happen to the equilibrium solution if just the prey is harvested? What if just the predator is harvested? What if both are harvested?
- c) Find the location of the equilibrium system in each of the three cases in the previous part. Do this in terms of the constants H_1 and H_2 for all three cases.

Exercise 5.3.13: The general competing species model has the form

$$\frac{dx}{dt} = x(\rho_1 - \gamma_1 y - M_1 x) \quad \frac{dy}{dt} = y(\rho_2 - \gamma_2 x - M_2 y)$$

where ρ indicates the growth rate, M is related to the carrying capacity, and γ is connected to the interaction term. Assume that this model is being used to represent species A and B of fish living in a pond at time t , which is initially stocked with both species of fish. We want to analyze the behavior of this equation under different sets of coefficients.

- a) If $\rho_2/\gamma_2 > \rho_1/M_1$ and $\rho_2/M_2 > \rho_1/\gamma_1$, show that the only equilibrium populations in the pond are no fish, no fish of species A, or no fish of species B. What happens for large values of t ?
- b) If $\rho_1/M_1 > \rho_2/\gamma_2$ and $\rho_1/\gamma_1 > \rho_2/M_2$, show that the only equilibrium populations in the pond are no fish, no fish of species A, or no fish of species B. What happens for large values of t ?
- c) Suppose that $\rho_2/\gamma_2 > \rho_1/M_1$ and $\rho_1/\gamma_1 > \rho_2/M_2$. Show that there is a stable equilibrium where both species coexist.

Exercise 5.3.14 (challenging): Take the pendulum, suppose the initial position is $\theta = 0$.

- a) Find the expression for ω giving the trajectory with initial condition $(0, \omega_0)$. Hint: Figure out what C should be in terms of ω_0 .
- b) Find the crucial angular velocity ω_1 , such that for any higher initial angular velocity, the pendulum will keep going around its axis, and for any lower initial angular velocity, the pendulum will simply swing back and forth. Hint: When the pendulum doesn't go over the top the expression for ω will be undefined for some θ s.
- c) What do you think happens if the initial condition is $(0, \omega_1)$, that is, the initial angle is 0, and the initial angular velocity is exactly ω_1 .

5.4 Limit cycles

Attribution: [JL], §8.4.

Learning Objectives

After this section, you will be able to:

- Identify differential equations that have limit cycles from slope fields and
- Find and classify limit cycles of systems of differential equations by converting the system to depend on radius.

For nonlinear systems, trajectories do not simply need to approach or leave a single point. They may in fact approach a larger set, such as a circle or another closed curve.

Example 5.4.1: The *Van der Pol oscillator** is the following equation

$$x'' - \mu(1 - x^2)x' + x = 0,$$

where μ is some positive constant. The Van der Pol oscillator originated with electrical circuits, but finds applications in diverse fields such as biology, seismology, and other physical sciences.

For simplicity, let us use $\mu = 1$. A phase diagram is given in the left-hand plot in Figure 5.11. Notice how the trajectories seem to very quickly settle on a closed curve. On the right-hand side is the plot of a single solution for $t = 0$ to $t = 30$ with initial conditions $x(0) = 0.1$ and $x'(0) = 0.1$. The solution quickly tends to a periodic solution.

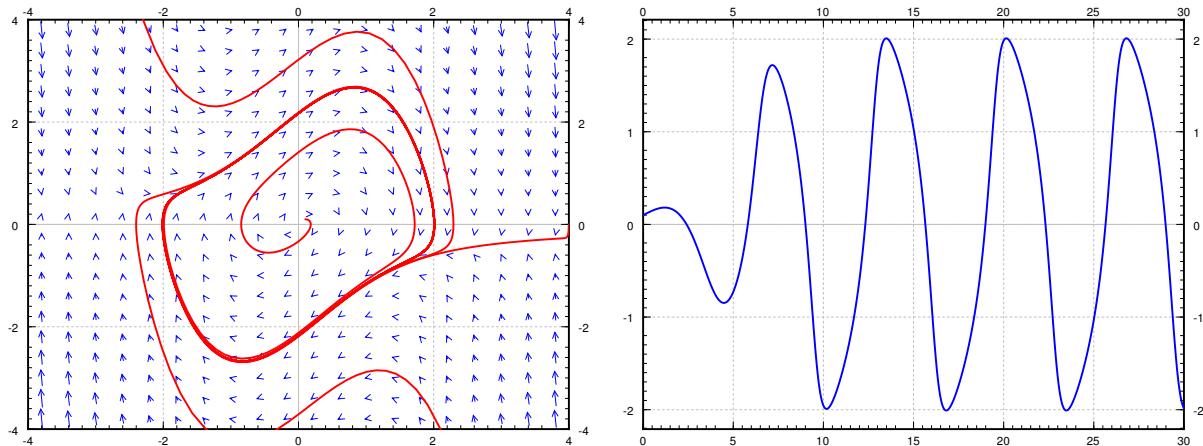


Figure 5.11: The phase portrait (left) and a graph of a sample solution of the Van der Pol oscillator.

The Van der Pol oscillator is an example of so-called *relaxation oscillation*. The word relaxation comes from the sudden jump (the very steep part of the solution). For larger μ

*Named for the Dutch physicist Balthasar van der Pol (1889–1959).

the steep part becomes even more pronounced, for small μ the limit cycle looks more like a circle. In fact, setting $\mu = 0$, we get $x'' + x = 0$, which is a linear system with a center and all trajectories become circles.

A trajectory in the phase portrait that is a closed curve (a curve that is a loop) is called a *closed trajectory*. A *limit cycle* is a closed trajectory such that at least one other trajectory spirals into it (or spirals out of it). For example, the closed curve in the phase portrait for the Van der Pol equation is a limit cycle. If all trajectories that start near the limit cycle spiral into it, the limit cycle is called *asymptotically stable*. The limit cycle in the Van der Pol oscillator is asymptotically stable.

Given a closed trajectory on an autonomous system, any solution that starts on it is periodic. Such a curve is called a *periodic orbit*. More precisely, if $(x(t), y(t))$ is a solution such that for some t_0 the point $(x(t_0), y(t_0))$ lies on a periodic orbit, then both $x(t)$ and $y(t)$ are periodic functions (with the same period). That is, there is some number P such that $x(t) = x(t + P)$ and $y(t) = y(t + P)$.

Consider the system

$$x' = f(x, y), \quad y' = g(x, y), \quad (5.4)$$

where the functions f and g have continuous derivatives in some region R in the plane.

Theorem 5.4.1 (Poincaré–Bendixson)

Suppose R is a closed bounded region (a region in the plane that includes its boundary and does not have points arbitrarily far from the origin). Suppose $(x(t), y(t))$ is a solution of (5.4) in R that exists for all $t \geq t_0$. Then either the solution is a periodic function, or the solution tends towards a periodic solution in R .

The main point of the theorem* is that if you find one solution that exists for all t large enough (that is, as t goes to infinity) and stays within a bounded region, then you have found either a periodic orbit, or a solution that spirals towards a limit cycle or tends to a critical point. That is, in the long term, the behavior is very close to a periodic function. Note that a constant solution at a critical point is periodic (with any period). The theorem is more a qualitative statement rather than something to help us in computations. In practice it is hard to find analytic solutions and so hard to show rigorously that they exist for all time. But if we think the solution exists we numerically solve for a large time to approximate the limit cycle. Another caveat is that the theorem only works in two dimensions. In three dimensions and higher, there is simply too much room.

The theorem applies to all solutions in the Van der Pol oscillator. Solutions that start at any point except the origin $(0, 0)$ will tend to the periodic solution around the limit cycle, and if the initial condition of $(0, 0)$ will lead to the constant solution $x = 0, y = 0$.

Example 5.4.2: Consider

$$x' = y + (x^2 + y^2 - 1)^2 x, \quad y' = -x + (x^2 + y^2 - 1)^2 y.$$

*Ivar Otto Bendixson (1861–1935) was a Swedish mathematician.

A vector field along with solutions with initial conditions $(1.02, 0)$, $(0.9, 0)$, and $(0.1, 0)$ are drawn in [Figure 5.12](#). Analyze this system to determine what will happen to the solution for a variety of initial conditions.

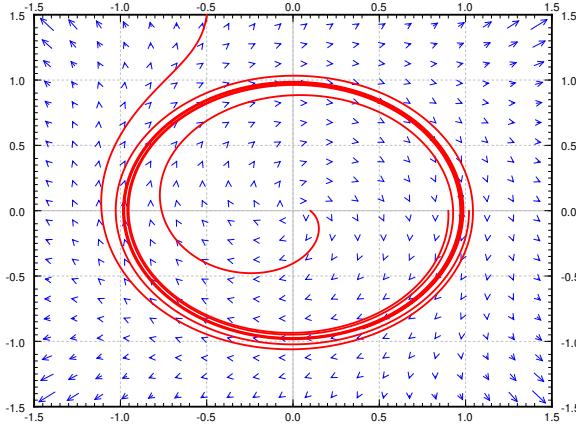


Figure 5.12: Unstable limit cycle example.

Solution: Notice that points on the unit circle (distance one from the origin) satisfy $x^2 + y^2 - 1 = 0$. And $x(t) = \sin(t)$, $y = \cos(t)$ is a solution of the system. Therefore we have a closed trajectory. For points off the unit circle, the second term in x' pushes the solution further away from the y -axis than the system $x' = y$, $y' = -x$, and y' pushes the solution further away from the x -axis than the linear system $x' = y$, $y' = -x$. In other words for all other initial conditions the trajectory will spiral out.

This means that for initial conditions inside the unit circle, the solution spirals out towards the periodic solution on the unit circle, and for initial conditions outside the unit circle the solutions spiral off towards infinity. Therefore the unit circle is a limit cycle, but not an asymptotically stable one. The Poincaré–Bendixson Theorem applies to the initial points inside the unit circle, as those solutions stay bounded, but not to those outside, as those solutions go off to infinity. □

A very similar analysis applies to the system

$$x' = y + (x^2 + y^2 - 1)x, \quad y' = -x + (x^2 + y^2 - 1)y.$$

We still obtain a closed trajectory on the unit circle, and points outside the unit circle spiral out to infinity, but now points inside the unit circle spiral towards the critical point at the origin. So this system does not have a limit cycle, even though it has a closed trajectory.

One way to see this more explicitly is by trying to write this all in terms of

$$r = \sqrt{x^2 + y^2}.$$

For simplicity here, we will determine everything in terms of

$$s = r^2 = x^2 + y^2$$

because as long as $r > 0$, r and s always have the same behavior (in terms of increasing and decreasing), and it is easier to compute with s .

Using the first example

$$x' = y + (x^2 + y^2 - 1)^2 x, \quad y' = -x + (x^2 + y^2 - 1)^2 y.$$

we see that

$$\begin{aligned} s' &= 2xx' + 2yy' \\ &= 2x(y + (x^2 + y^2 - 1)^2 x) + 2y(-x + (x^2 + y^2 - 1)^2 y) \\ &= 2xy + 2x^2(x^2 + y^2 - 1)^2 - 2xy + 2y^2(x^2 + y^2 - 1)^2 \\ s' &= 2s(s - 1)^2 \end{aligned}$$

Thus, we are left with the equation

$$\frac{ds}{dt} = 2s(s - 1)^2$$

which is an autonomous first-order equation that we can analyze. We have two equilibrium solutions in terms of s at $s = 0$, which corresponds to the origin, and $s = 1$, which corresponds to the unit circle. We can then plug in values to see that for $s = \frac{1}{2}$, $\frac{ds}{dt} > 0$, so that the solutions will increase out to the unit circle. For $s > 1$, $\frac{ds}{dt} > 0$ as well, so solutions move away from the circle outside it. This is the same as the result we obtained in the first example.

For the second example, we end up with the autonomous equation

$$\frac{ds}{dt} = 2s(s - 1)$$

which is negative for $0 < s < 1$ and positive for $1 < s$, giving that solutions that start inside the unit circle will converge to the origin, and solutions that start outside the circle will move away from it.

Due to the Picard theorem ([Theorem 4.1.1](#) on page 244) we find that no matter where we are in the plane we can always find a solution a little bit further in time, as long as f and g have continuous derivatives. So if we find a closed trajectory in an autonomous system, then for every initial point inside the closed trajectory, the solution will exist for all time and it will stay bounded (it will stay inside the closed trajectory). So the moment we found the solution above going around the unit circle, we knew that for every initial point inside the circle, the solution exists for all time and the Poincaré–Bendixson theorem applies.

Let us next look for conditions when limit cycles (or periodic orbits) do not exist. We assume the equation (5.4) is defined on a *simply connected region*, that is, a region with no holes we can go around. For example the entire plane is a simply connected region, and so is the inside of the unit disc. However, the entire plane minus a point is not a simply connected domain as it has a “hole” at the origin.

Theorem 5.4.2 (Bendixson–Dulac)

Suppose R is a simply connected region, and the expression^a

$$\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y}$$

is either always positive or always negative on R (except perhaps a small set such as on isolated points or curves) then the system (5.4) has no closed trajectory inside R .

^aUsually the expression in the Bendixson–Dulac Theorem is $\frac{\partial(\varphi f)}{\partial x} + \frac{\partial(\varphi g)}{\partial y}$ for some continuously differentiable function φ . For simplicity, let us just consider the case $\varphi = 1$.

The theorem* gives us a way of ruling out the existence of a closed trajectory, and hence a way of ruling out limit cycles. The exception about points or curves means that we can allow the expression to be zero at a few points, or perhaps on a curve, but not on any larger set.

Example 5.4.3: Let us look at $x' = y + y^2 e^x$, $y' = x$ in the entire plane (see [Example 5.2.2](#) on page 322). The entire plane is simply connected and so we can apply the theorem. We compute $\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = y^2 e^x + 0$. The function $y^2 e^x$ is always positive except on the line $y = 0$. Therefore, via the theorem, the system has no closed trajectories.

In some books (or the internet) the theorem is not stated carefully and it concludes there are no periodic solutions. That is not quite right. The example above has two critical points and hence it has constant solutions, and constant functions are periodic. The conclusion of the theorem should be that there exist no trajectories that form closed curves. Another way to state the conclusion of the theorem would be to say that there exist no nonconstant periodic solutions that stay in R .

Example 5.4.4: Let us look at a somewhat more complicated example. Take the system $x' = -y - x^2$, $y' = -x + y^2$ (see [Example 5.2.1](#) on page 321). We compute $\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} = -2x + 2y = 2(-x + y)$. This expression takes on both signs, so if we are talking about the whole plane we cannot simply apply the theorem. However, we could apply it on the set where $-x + y \geq 0$. Via the theorem, there is no closed trajectory in that set. Similarly, there is no closed trajectory in the set $-x + y \leq 0$. We cannot conclude (yet) that there is no closed trajectory in the entire plane. Perhaps half of it is in the set where $-x + y \geq 0$ and the other half is in the set where $-x + y \leq 0$.

The key is to look at the line where $-x + y = 0$, or $x = y$. On this line $x' = -y - x^2 = -x - x^2$ and $y' = -x + y^2 = -x + x^2$. In particular, when $x = y$ then $x' \leq y'$. That means that the arrows, the vectors (x', y') , always point into the set where $-x + y \geq 0$. There is no way we can start in the set where $-x + y \geq 0$ and go into the set where $-x + y \leq 0$. Once we are in the set where $-x + y \geq 0$, we stay there. So no closed trajectory can have points in both sets.

Example 5.4.5: Consider $x' = y + (x^2 + y^2 - 1)x$, $y' = -x + (x^2 + y^2 - 1)y$, and consider the region R given by $x^2 + y^2 > \frac{1}{2}$. That is, R is the region outside a circle of radius $\frac{1}{\sqrt{2}}$

*Henri Dulac (1870–1955) was a French mathematician.

centered at the origin. Then there is a closed trajectory in R , namely $x = \cos(t)$, $y = \sin(t)$. Furthermore,

$$\frac{\partial f}{\partial x} + \frac{\partial g}{\partial x} = 4x^2 + 4y^2 - 2,$$

which is always positive on R . So what is going on? The Bendixson–Dulac theorem does not apply since the region R is not simply connected—it has a hole, the circle we cut out!

5.4.1 Exercises

Exercise 5.4.1: Consider the two-dimensional system of differential equation written in polar coordinates as

$$\frac{dr}{dt} = r(r-1)(r-4)^2 \quad \frac{d\theta}{dt} = 1.$$

Determine all limit cycles, periodic solutions, and classify the stability of each of these solutions.

Exercise 5.4.2: Consider the two-dimensional system of differential equation written in polar coordinates as

$$\frac{dr}{dt} = r^2(r-1)^2(r-3) \quad \frac{d\theta}{dt} = -1.$$

Determine all limit cycles, periodic solutions, and classify the stability of each of these solutions.

Exercise 5.4.3:* Consider the system of differential equation given by

$$\frac{dx}{dt} = x(3 - 2y^2 - x^2) \quad \frac{dy}{dt} = y(3 - y^2).$$

Find and classify all limit cycles by converting to an autonomous equation in $r = \sqrt{x^2 + y^2}$ or $s = x^2 + y^2$.

Exercise 5.4.4:* Consider the system of differential equation given by

$$\frac{dx}{dt} = -x(x^2 + y^2)^2 + 6x(x^2 + y^2) - 8x + 6y \quad \frac{dy}{dt} = -y(x^2 + y^2)^2 + 6y(x^2 + y^2) - 8y - 6x.$$

Find and classify all limit cycles by converting to an autonomous equation in $r = \sqrt{x^2 + y^2}$ or $s = x^2 + y^2$.

Exercise 5.4.5: Show that the following systems have no closed trajectories.

a) $x' = x^3 + y$, $y' = y^3 + x^2$,

b) $x' = e^{x-y}$, $y' = e^{x+y}$,

c) $x' = x + 3y^2 - y^3$, $y' = y^3 + x^2$.

Exercise 5.4.6:* Show that the following systems have no closed trajectories.

a) $x' = x + y^2$, $y' = y + x^2$,

b) $x' = -x \sin^2(y)$, $y' = e^x$,

c) $x' = xy$, $y' = x + x^2$.

Exercise 5.4.7:* Suppose an autonomous system in the plane has a solution $x = \cos(t) + e^{-t}$, $y = \sin(t) + e^{-t}$. What can you say about the system (in particular about limit cycles and periodic solutions)?

Exercise 5.4.8: Formulate a condition for a 2-by-2 linear system $\vec{x}' = A\vec{x}$ to not be a center using the Bendixson–Dulac theorem. That is, the theorem says something about certain elements of A .

Exercise 5.4.9: Explain why the Bendixson–Dulac Theorem does not apply for any conservative system $x'' + h(x) = 0$.

Exercise 5.4.10: A system such as $x' = x$, $y' = y$ has solutions that exist for all time t , yet there are no closed trajectories. Explain why the Poincaré–Bendixson Theorem does not apply.

Exercise 5.4.11:* Show that the limit cycle of the Van der Pol oscillator (for $\mu > 0$) must not lie completely in the set where $-1 < x < 1$. Compare with [Figure 5.11](#) on page 343.

Exercise 5.4.12: Differential equations can also be given in different coordinate systems. Suppose we have the system $r' = 1 - r^2$, $\theta' = 1$ given in polar coordinates. Find all the closed trajectories and check if they are limit cycles and if so, if they are asymptotically stable or not.

Exercise 5.4.13:* Suppose we have the system $r' = \sin(r)$, $\theta' = 1$ given in polar coordinates. Find all the closed trajectories.

5.5 Chaos

Attribution: [JL], §8.5.

Learning Objectives

After this section, you will be able to:

- Identify chaotic behavior and how it is distinct from other types of equations.

You have surely heard the story about the flap of a butterfly wing in the Amazon causing hurricanes in the North Atlantic. In a prior section, we mentioned that a small change in initial conditions of the planets can lead to very different configuration of the planets in the long term. These are examples of *chaotic systems*. Mathematical chaos is not really chaos, there is precise order behind the scenes. Everything is still deterministic. However a chaotic system is extremely sensitive to initial conditions. This also means even small errors induced via numerical approximation create large errors very quickly, so it is almost impossible to numerically approximate for long times. This is a large part of the trouble, as chaotic systems cannot be in general solved analytically.

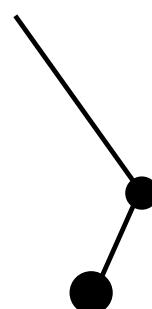
Take the weather, the most well-known chaotic system. A small change in the initial conditions (the temperature at every point of the atmosphere for example) produces drastically different predictions in relatively short time, and so we cannot accurately predict weather. And we do not actually know the exact initial conditions. We measure temperatures at a few points with some error, and then we somehow estimate what is in between. There is no way we can accurately measure the effects of every butterfly wing. Then we solve the equations numerically introducing new errors. You should not trust weather prediction more than a few days out.

Chaotic behavior was first noticed by Edward Lorenz* in the 1960s when trying to model thermally induced air convection (movement). Lorenz was looking at the relatively simple system:

$$x' = -10x + 10y, \quad y' = 28x - y - xz, \quad z' = -\frac{8}{3}z + xy.$$

A small change in the initial conditions yields a very different solution after a reasonably short time.

A simple example the reader can experiment with, and which displays chaotic behavior, is a double pendulum. The equations for this setup are somewhat complicated, and their derivation is quite tedious, so we will not bother to write them down. The idea is to put a pendulum on the end of another pendulum. The movement of the bottom mass will appear chaotic. This type of chaotic system is a basis for a whole number of office novelty desk toys. It is simple to build a version. Take a piece of a string. Tie two heavy nuts at different points of the string; one at the end, and one a bit above. Now give the bottom nut a little push. As long as the swings are not too big and the string stays tight, you have a double pendulum system.



*Edward Norton Lorenz (1917–2008) was an American mathematician and meteorologist.

5.5.1 Duffing equation and strange attractors

Let us study the so-called *Duffing equation*:

$$x'' + ax' + bx + cx^3 = C \cos(\omega t).$$

Here a , b , c , C , and ω are constants. Except for the cx^3 term, this equation looks like a forced mass-spring system. The cx^3 means the spring does not exactly obey Hooke's law (which no real-world spring actually does obey exactly). When c is not zero, the equation does not have a closed form solution, so we must resort to numerical solutions, as is usual for nonlinear systems. Not all choices of constants and initial conditions exhibit chaotic behavior. Let us study

$$x'' + 0.05x' + x^3 = 8 \cos(t).$$

The equation is not autonomous, so we cannot draw the vector field in the phase plane. We can still draw the trajectories. In Figure 5.13 we plot trajectories for t going from 0 to 15, for two very close initial conditions $(2, 3)$ and $(2, 2.9)$, and also the solutions in the (x, t) space. The two trajectories are close at first, but after a while diverge significantly. This sensitivity to initial conditions is precisely what we mean by the system behaving chaotically.

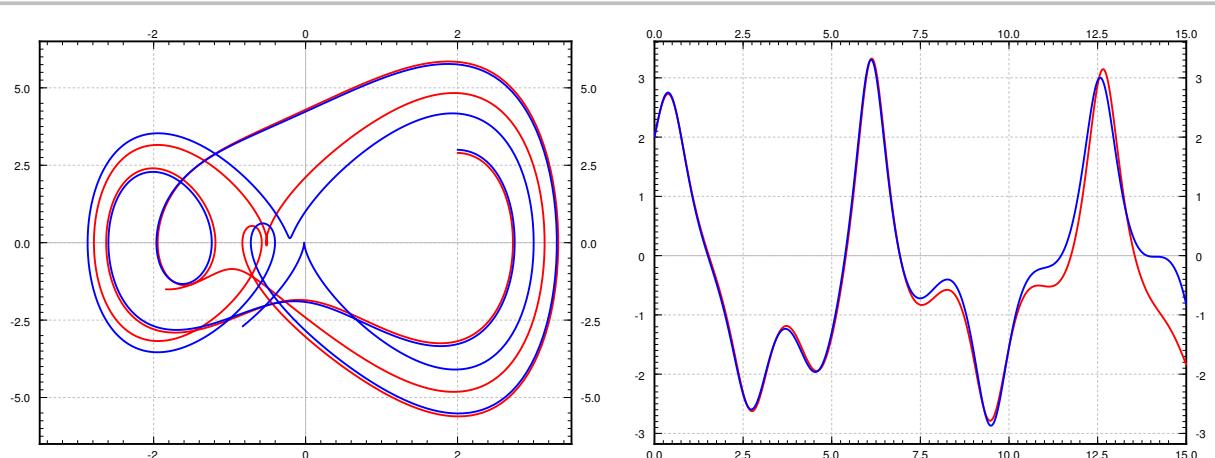


Figure 5.13: On left, two trajectories in phase space for $0 \leq t \leq 15$, for the Duffing equation one with initial conditions $(2, 3)$ and the other with $(2, 2.9)$. On right the two solutions in (x, t) -space.

Let us see the long term behavior. In Figure 5.14 on the following page, we plot the behavior of the system for initial conditions $(2, 3)$ for a longer period of time. It is hard to see any particular pattern in the shape of the solution except that it seems to oscillate, but each oscillation appears quite unique. The oscillation is expected due to the forcing term. We mention that to produce the picture accurately, a ridiculously large number of steps* had to be used in the numerical algorithm, as even small errors quickly propagate in a chaotic system.

*In fact for reference, 30,000 steps were used with the Runge–Kutta algorithm, see exercises in § 1.6.

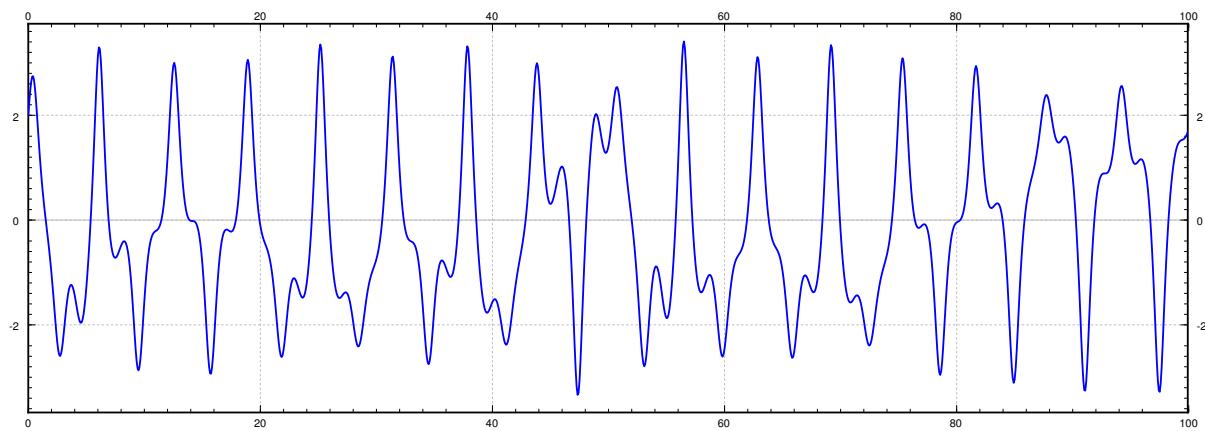


Figure 5.14: The solution to the given Duffing equation for t from 0 to 100.

It is very difficult to analyze chaotic systems, or to find the order behind the madness, but let us try to do something that we did for the standard mass-spring system. One way we analyzed the system is that we figured out what was the long term behavior (not dependent on initial conditions). From the figure above, it is clear that we will not get a nice exact description of the long term behavior for this chaotic system, but perhaps we can find some order to what happens on each “oscillation” and what do these oscillations have in common.

The concept we explore is that of a *Poincaré section*^{*}. Instead of looking at t in a certain interval, we look at where the system is at a certain sequence of points in time. Imagine flashing a strobe at a fixed frequency and drawing the points where the solution is during the flashes. The right strobing frequency depends on the system in question. The correct frequency for the forced Duffing equation (and other similar systems) is the frequency of the forcing term. For the Duffing equation above, find a solution $(x(t), y(t))$, and look at the points

$$(x(0), y(0)), \quad (x(2\pi), y(2\pi)), \quad (x(4\pi), y(4\pi)), \quad (x(6\pi), y(6\pi)), \quad \dots$$

As we are really not interested in the transient part of the solution, that is, the part of the solution that depends on the initial condition, we skip some number of steps in the beginning. For example, we might skip the first 100 such steps and start plotting points at $t = 100(2\pi)$, that is

$$(x(200\pi), y(200\pi)), \quad (x(202\pi), y(202\pi)), \quad (x(204\pi), y(204\pi)), \quad \dots$$

The plot of these points is the Poincaré section. After plotting enough points, a curious pattern emerges in Figure 5.15 on the next page (the left-hand picture), a so-called *strange attractor*.

Given a sequence of points, an *attractor* is a set towards which the points in the sequence eventually get closer and closer to, that is, they are attracted. The Poincaré section is not really the attractor itself, but as the points are very close to it, we see its shape. The strange

*Named for the French polymath Jules Henri Poincaré (1854–1912).

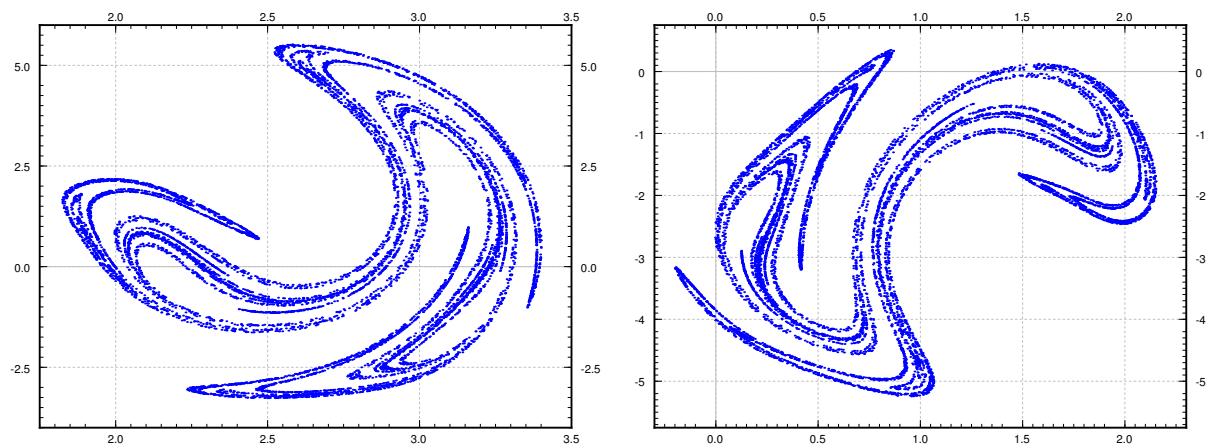


Figure 5.15: Strange attractor. The left plot is with no phase shift, the right plot has phase shift $\pi/4$.

attractor is a very complicated set. It has fractal structure, that is, if you zoom in as far as you want, you keep seeing the same complicated structure.

The initial condition makes no difference. If we start with a different initial condition, the points eventually gravitate towards the attractor, and so as long as we throw away the first few points, we get the same picture. Similarly small errors in the numerical approximations do not matter here.

An amazing thing is that a chaotic system such as the Duffing equation is not random at all. There is a very complicated order to it, and the strange attractor says something about this order. We cannot quite say what state the system will be in eventually, but given the fixed strobing frequency we narrow it down to the points on the attractor.

If we use a phase shift, for example $\pi/4$, and look at the times

$$\pi/4, \quad 2\pi + \pi/4, \quad 4\pi + \pi/4, \quad 6\pi + \pi/4, \quad \dots$$

we obtain a slightly different attractor. The picture is the right-hand side of Figure 5.15. It is as if we had rotated, moved, and slightly distorted the original. For each phase shift you can find the set of points towards which the system periodically keeps coming back to.

Study the pictures and notice especially the scales—where are these attractors located in the phase plane. Notice the regions where the strange attractor lives and compare it to the plot of the trajectories in Figure 5.13 on page 351.

Let us compare this section to the discussion in § 2.6 about forced oscillations. Take the equation

$$x'' + 2px' + \omega_0^2 x = \frac{F_0}{m} \cos(\omega t).$$

This is like the Duffing equation, but with no x^3 term. The steady periodic solution is of the form

$$x = C \cos(\omega t + \gamma).$$

Strobing using the frequency ω , we obtain a single point in the phase space. The attractor in this setting is a single point—an expected result as the system is not chaotic. It was the opposite of chaotic: Any difference induced by the initial conditions dies away very quickly, and we settle into always the same steady periodic motion.

5.5.2 The Lorenz system

In two dimensions to find chaotic behavior, we must study forced, or non-autonomous, systems such as the Duffing equation. The Poincaré–Bendixson Theorem says that a solution to an autonomous two-dimensional system that exists for all time in the future and does not go towards infinity is periodic or tends towards a periodic solution. Hardly the chaotic behavior we are looking for.

In three dimensions even autonomous systems can be chaotic. Let us very briefly return to the Lorenz system

$$x' = -10x + 10y, \quad y' = 28x - y - xz, \quad z' = -\frac{8}{3}z + xy.$$

The Lorenz system is an autonomous system in three dimensions exhibiting chaotic behavior. See the [Figure 5.16](#) for a sample trajectory, which is now a curve in three-dimensional space.

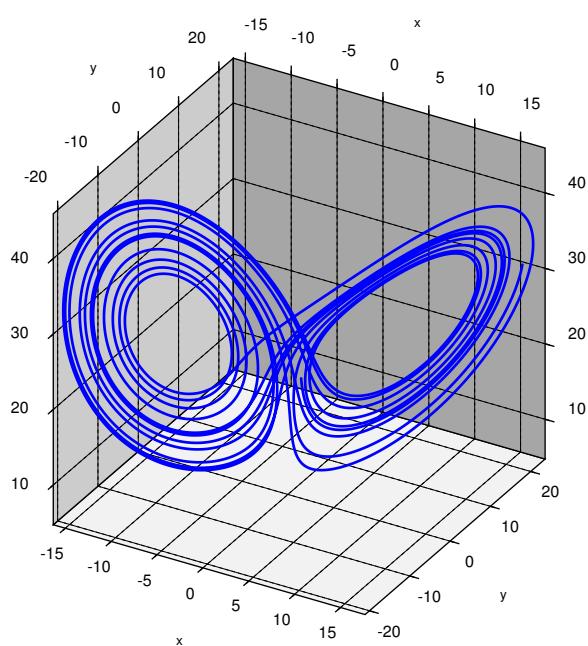


Figure 5.16: A trajectory in the Lorenz system.

The solutions tend to an *attractor* in space, the so-called *Lorenz attractor*. In this case no strobing is necessary. Again we cannot quite see the attractor itself, but if we try to follow a solution for long enough, as in the figure, we get a pretty good picture of what the attractor looks like. The Lorenz attractor is also a strange attractor and has a complicated fractal

structure. And, just as for the Duffing equation, what we want to draw is not the whole trajectory, but start drawing the trajectory after a while, once it is close to the attractor.

The path of the trajectory is not simply a repeating figure-eight. The trajectory spins some seemingly random number of times on the left, then spins a number of times on the right, and so on. As this system arose in weather prediction, one can perhaps imagine a few days of warm weather and then a few days of cold weather, where it is not easy to predict when the weather will change, just as it is not really easy to predict far in advance when the solution will jump onto the other side. See [Figure 5.17](#) for a plot of the x component of the solution drawn above. A negative x corresponds to the left “loop” and a positive x corresponds to the right “loop”.

Most of the mathematics we studied in this book is quite classical and well understood. On the other hand, chaos, including the Lorenz system, continues to be the subject of current research. Furthermore, chaos has found applications not just in the sciences, but also in art.

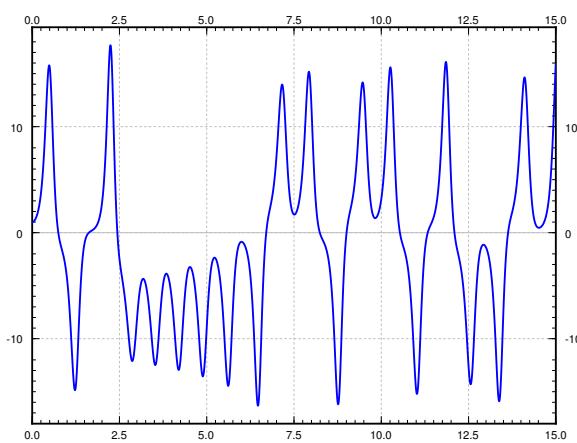


Figure 5.17: Graph of the $x(t)$ component of the solution.

5.5.3 Exercises

Exercise 5.5.1 (*): Find critical points of the Lorenz system and the associated linearizations.

Exercise 5.5.2: For the non-chaotic equation $x'' + 2px' + \omega_0^2 x = \frac{F_0}{m} \cos(\omega t)$, suppose we strobe with frequency ω as we mentioned above. Use the known steady periodic solution to find precisely the point which is the attractor for the Poincaré section.

Exercise 5.5.3 (project): Construct the double pendulum described in the text with a string and two nuts (or heavy beads). Play around with the position of the middle nut, and perhaps use different weight nuts. Describe what you find.

Exercise 5.5.4 (project): A simple fractal attractor can be drawn via the following chaos game. Draw the three vertices of a triangle and label them, say p_1 , p_2 and p_3 . Draw some random point p (it does not have to be one of the three points above). Roll a die to pick of the p_1 , p_2 , or p_3 randomly (for example 1 and 4 mean p_1 , 2 and 5 mean p_2 , and 3 and 6 mean p_3). Suppose we picked p_2 , then let p_{new} be the point exactly halfway between p and p_2 . Draw this point and let p now refer to this new point p_{new} . Rinse, repeat. Try to be precise and draw as many iterations as possible. Your points will be attracted to the so-called Sierpinski triangle. A computer was used to run the game for 10,000 iterations to obtain the picture in [Figure 5.18](#).

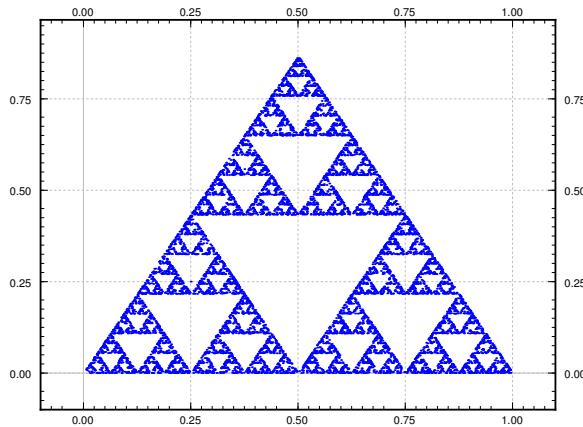


Figure 5.18: 10,000 iterations of the chaos game producing the Sierpinski triangle.

Exercise 5.5.5 (computer project): Use a computer software (such as Matlab, Octave, or perhaps even a spreadsheet), plot the solution of the given forced Duffing equation with Euler's method. Plotting the solution for t from 0 to 100 with several different (small) step sizes. Discuss.

Chapter 6

The Laplace transform

6.1 The Laplace transform

Attribution: [JL], §6.1.

Learning Objectives

After this section, you will be able to:

- Define the Laplace transform,
- Compute the Laplace transform of a variety of functions, and
- Compute the inverse Laplace transform of functions when it exists.

6.1.1 The transform

In this chapter we will discuss the Laplace transform[†]. The Laplace transform is a very efficient method to solve certain ODE or PDE problems. The transform takes a differential equation and turns it into an algebraic equation. If the algebraic equation can be solved, applying the inverse transform gives us our desired solution. The Laplace transform also has applications in the analysis of electrical circuits, NMR spectroscopy, signal processing, and elsewhere. Finally, understanding the Laplace transform will also help with understanding the related Fourier transform, which, however, requires more understanding of complex numbers. We will not cover the Fourier transform.

The Laplace transform also gives a lot of insight into the nature of the equations we are dealing with. It can be seen as converting between the time and the frequency domain. For example, take the standard equation

$$mx''(t) + cx'(t) + kx(t) = f(t).$$

[†]Just like the Laplace equation and the Laplacian, the Laplace transform is also named after Pierre-Simon, marquis de Laplace (1749–1827).

We can think of t as time and $f(t)$ as incoming signal. The Laplace transform will convert the equation from a differential equation in time to an algebraic (no derivatives) equation, where the new independent variable s is the frequency.

We can think of the *Laplace transform* as a black box. It eats functions and spits out functions in a new variable. We write $\mathcal{L}\{f(t)\} = F(s)$ for the Laplace transform of $f(t)$. It is common to write lower case letters for functions in the time domain and upper case letters for functions in the frequency domain. We use the same letter to denote that one function is the Laplace transform of the other. For example, $F(s)$ is the Laplace transform of $f(t)$. Let us define the transform.

$$\mathcal{L}\{f(t)\} = F(s) \stackrel{\text{def}}{=} \int_0^\infty e^{-st} f(t) dt.$$

We note that we are only considering $t \geq 0$ in the transform. Of course, if we think of t as time there is no problem, we are generally interested in finding out what will happen in the future (Laplace transform is one place where it is safe to ignore the past). Let us compute some simple transforms.

Example 6.1.1: Suppose $f(t) = 1$, then

$$\mathcal{L}\{1\} = \int_0^\infty e^{-st} dt = \left[\frac{e^{-st}}{-s} \right]_{t=0}^\infty = \lim_{h \rightarrow \infty} \left[\frac{e^{-st}}{-s} \right]_{t=0}^h = \lim_{h \rightarrow \infty} \left(\frac{e^{-sh}}{-s} - \frac{1}{-s} \right) = \frac{1}{s}.$$

The limit (the improper integral) only exists if $s > 0$. So $\mathcal{L}\{1\}$ is only defined for $s > 0$.

Example 6.1.2: Suppose $f(t) = e^{-at}$, then

$$\mathcal{L}\{e^{-at}\} = \int_0^\infty e^{-st} e^{-at} dt = \int_0^\infty e^{-(s+a)t} dt = \left[\frac{e^{-(s+a)t}}{-(s+a)} \right]_{t=0}^\infty = \frac{1}{s+a}.$$

The limit only exists if $s + a > 0$. So $\mathcal{L}\{e^{-at}\}$ is only defined for $s + a > 0$.

Example 6.1.3: Suppose $f(t) = t$, then using integration by parts

$$\begin{aligned} \mathcal{L}\{t\} &= \int_0^\infty e^{-st} t dt \\ &= \left[\frac{-te^{-st}}{s} \right]_{t=0}^\infty + \frac{1}{s} \int_0^\infty e^{-st} dt \\ &= 0 + \frac{1}{s} \left[\frac{e^{-st}}{-s} \right]_{t=0}^\infty \\ &= \frac{1}{s^2}. \end{aligned}$$

Again, the limit only exists if $s > 0$.

Example 6.1.4: A common function is the *unit step function*, which is sometimes called the *Heaviside function**. This function is generally given as

$$u(t) = \begin{cases} 0 & \text{if } t < 0, \\ 1 & \text{if } t \geq 0. \end{cases}$$

*The function is named after the English mathematician, engineer, and physicist Oliver Heaviside (1850–1925). Only by coincidence is the function “heavy” on “one side.”

Let us find the Laplace transform of $u(t - a)$, where $a \geq 0$ is some constant. That is, the function that is 0 for $t < a$ and 1 for $t \geq a$.

$$\mathcal{L}\{u(t - a)\} = \int_0^\infty e^{-st} u(t - a) dt = \int_a^\infty e^{-st} dt = \left[\frac{e^{-st}}{-s} \right]_{t=a}^\infty = \frac{e^{-as}}{s},$$

where of course $s > 0$ (and $a \geq 0$ as we said before).

By applying similar procedures we can compute the transforms of many elementary functions. Many basic transforms are listed in [Table 6.1](#).

$f(t)$	$\mathcal{L}\{f(t)\}$	$f(t)$	$\mathcal{L}\{f(t)\}$
C	$\frac{C}{s}$	$\sin(\omega t)$	$\frac{\omega}{s^2 + \omega^2}$
t	$\frac{1}{s^2}$	$\cos(\omega t)$	$\frac{s}{s^2 + \omega^2}$
t^2	$\frac{2}{s^3}$	$\sinh(\omega t)$	$\frac{\omega}{s^2 - \omega^2}$
t^3	$\frac{6}{s^4}$	$\cosh(\omega t)$	$\frac{s}{s^2 - \omega^2}$
t^n	$\frac{n!}{s^{n+1}}$	$u(t - a)$	$\frac{e^{-as}}{s}$
e^{-at}	$\frac{1}{s+a}$		

Table 6.1: Some Laplace transforms (C , ω , and a are constants).

Exercise 6.1.1: Verify [Table 6.1](#).

Since the transform is defined by an integral. We can use the linearity properties of the integral. For example, suppose C is a constant, then

$$\mathcal{L}\{Cf(t)\} = \int_0^\infty e^{-st} Cf(t) dt = C \int_0^\infty e^{-st} f(t) dt = C\mathcal{L}\{f(t)\}.$$

So we can “pull out” a constant out of the transform. Similarly we have linearity. Since linearity is very important we state it as a theorem.

Theorem 6.1.1 (Linearity of the Laplace transform)

Suppose that A , B , and C are constants, then

$$\mathcal{L}\{Af(t) + Bg(t)\} = A\mathcal{L}\{f(t)\} + B\mathcal{L}\{g(t)\},$$

and in particular

$$\mathcal{L}\{Cf(t)\} = C\mathcal{L}\{f(t)\}.$$

Exercise 6.1.2: Verify the theorem. That is, show that $\mathcal{L}\{Af(t) + Bg(t)\} = A\mathcal{L}\{f(t)\} + B\mathcal{L}\{g(t)\}$.

These rules together with Table 6.1 on the previous page make it easy to find the Laplace transform of a whole lot of functions already. But be careful. It is a common mistake to think that the Laplace transform of a product is the product of the transforms. In general

$$\mathcal{L}\{f(t)g(t)\} \neq \mathcal{L}\{f(t)\}\mathcal{L}\{g(t)\}.$$

It must also be noted that not all functions have a Laplace transform. For example, the function $\frac{1}{t}$ does not have a Laplace transform as the integral diverges for all s . Similarly, $\tan t$ or e^{t^2} do not have Laplace transforms.

6.1.2 Existence and uniqueness

When does the Laplace transform exist? A function $f(t)$ is of *exponential order* as t goes to infinity if

$$|f(t)| \leq M e^{ct},$$

for some constants M and c , for sufficiently large t (say for all $t > t_0$ for some t_0). The simplest way to check this condition is to try and compute

$$\lim_{t \rightarrow \infty} \frac{f(t)}{e^{ct}}.$$

If the limit exists and is finite (usually zero), then $f(t)$ is of exponential order.

Exercise 6.1.3: Use L'Hopital's rule from calculus to show that a polynomial is of exponential order. Hint: Note that a sum of two exponential order functions is also of exponential order. Then show that t^n is of exponential order for any n .

For an exponential order function we have existence and uniqueness of the Laplace transform.

Theorem 6.1.2 (Existence)

Let $f(t)$ be continuous and of exponential order for a certain constant c . Then $F(s) = \mathcal{L}\{f(t)\}$ is defined for all $s > c$.

The existence is not difficult to see. Let $f(t)$ be of exponential order, that is $|f(t)| \leq M e^{ct}$ for all $t > 0$ (for simplicity $t_0 = 0$). Let $s > c$, or in other words $(c-s) < 0$. By the comparison theorem from calculus, the improper integral defining $\mathcal{L}\{f(t)\}$ exists if the following integral exists

$$\int_0^\infty e^{-st} (M e^{ct}) dt = M \int_0^\infty e^{(c-s)t} dt = M \left[\frac{e^{(c-s)t}}{c-s} \right]_{t=0}^\infty = \frac{M}{c-s}.$$

The transform also exists for some other functions that are not of exponential order, but that will not be relevant to us. Before dealing with uniqueness, let us note that for exponential order functions we obtain that their Laplace transform decays at infinity:

$$\lim_{s \rightarrow \infty} F(s) = 0.$$

Theorem 6.1.3 (Uniqueness)

Let $f(t)$ and $g(t)$ be continuous and of exponential order. Suppose that there exists a constant C , such that $F(s) = G(s)$ for all $s > C$. Then $f(t) = g(t)$ for all $t \geq 0$.

Both theorems hold for piecewise continuous functions as well. Recall that piecewise continuous means that the function is continuous except perhaps at a discrete set of points, where it has jump discontinuities like the Heaviside function. Uniqueness, however, does not “see” values at the discontinuities. So we can only conclude that $f(t) = g(t)$ outside of discontinuities. For example, the unit step function is sometimes defined using $u(0) = 1/2$. This new step function, however, has the exact same Laplace transform as the one we defined earlier where $u(0) = 1$.

6.1.3 The inverse transform

As we said, the Laplace transform will allow us to convert a differential equation into an algebraic equation. Once we solve the algebraic equation in the frequency domain we will want to get back to the time domain, as that is what we are interested in. Given a function $F(s)$, we wish to find a function $f(t)$ such that $\mathcal{L}\{f(t)\} = F(s)$. **Theorem 6.1.3** says that the solution $f(t)$ is unique. So we can without fear make the following definition.

Suppose $F(s) = \mathcal{L}\{f(t)\}$ for some function $f(t)$. Define the *inverse Laplace transform* as

$$\mathcal{L}^{-1}\{F(s)\} \stackrel{\text{def}}{=} f(t).$$

There is an integral formula for the inverse, but it is not as simple as the transform itself—it requires complex numbers and path integrals. For us it will suffice to compute the inverse using [Table 6.1](#) on page 359.

Example 6.1.5: Take $F(s) = \frac{1}{s+1}$. Find the inverse Laplace transform.

We look at the table to find

$$\mathcal{L}^{-1}\left\{\frac{1}{s+1}\right\} = e^{-t}.$$

As the Laplace transform is linear, the inverse Laplace transform is also linear. That is,

$$\mathcal{L}^{-1}\{AF(s) + BG(s)\} = A\mathcal{L}^{-1}\{F(s)\} + B\mathcal{L}^{-1}\{G(s)\}.$$

Of course, we also have $\mathcal{L}^{-1}\{AF(s)\} = A\mathcal{L}^{-1}\{F(s)\}$. Let us demonstrate how linearity can be used.

Example 6.1.6: Take $F(s) = \frac{s^2+s+1}{s^3+s}$. Find the inverse Laplace transform.

Solution: First we use the *method of partial fractions* to write F in a form where we can use [Table 6.1](#) on page 359. We factor the denominator as $s(s^2 + 1)$ and write

$$\frac{s^2+s+1}{s^3+s} = \frac{A}{s} + \frac{Bs+C}{s^2+1}.$$

Putting the right-hand side over a common denominator and equating the numerators we get $A(s^2 + 1) + s(Bs + C) = s^2 + s + 1$. Expanding and equating coefficients we obtain $A + B = 1$, $C = 1$, $A = 1$, and thus $B = 0$. In other words,

$$F(s) = \frac{s^2 + s + 1}{s^3 + s} = \frac{1}{s} + \frac{1}{s^2 + 1}.$$

By linearity of the inverse Laplace transform we get

$$\mathcal{L}^{-1}\left\{\frac{s^2 + s + 1}{s^3 + s}\right\} = \mathcal{L}^{-1}\left\{\frac{1}{s}\right\} + \mathcal{L}^{-1}\left\{\frac{1}{s^2 + 1}\right\} = 1 + \sin t.$$

□

Another useful property is the so-called *shifting property* or the *first shifting property*

$$\mathcal{L}\{e^{-at}f(t)\} = F(s + a),$$

where $F(s)$ is the Laplace transform of $f(t)$.

Exercise 6.1.4: Derive the first shifting property from the definition of the Laplace transform.

The shifting property can be used, for example, when the denominator is a more complicated quadratic that may come up in the method of partial fractions. We complete the square and write such quadratics as $(s + a)^2 + b$ and then use the shifting property.

Example 6.1.7: Find $\mathcal{L}^{-1}\left\{\frac{1}{s^2+4s+8}\right\}$.

Solution: First we complete the square to make the denominator $(s + 2)^2 + 4$. Next we find

$$\mathcal{L}^{-1}\left\{\frac{1}{s^2 + 4}\right\} = \frac{1}{2} \sin(2t).$$

Putting it all together with the shifting property, we find

$$\mathcal{L}^{-1}\left\{\frac{1}{s^2 + 4s + 8}\right\} = \mathcal{L}^{-1}\left\{\frac{1}{(s + 2)^2 + 4}\right\} = \frac{1}{2} e^{-2t} \sin(2t).$$

□

In general, we want to be able to apply the Laplace transform to rational functions, that is functions of the form

$$\frac{F(s)}{G(s)}$$

where $F(s)$ and $G(s)$ are polynomials. Since normally, for the functions that we are considering, the Laplace transform goes to zero as $s \rightarrow \infty$, it is not hard to see that the degree of $F(s)$ must be smaller than that of $G(s)$. Such rational functions are called *proper rational functions* and we can always apply the method of partial fractions. Of course this means we need to be able to factor the denominator into linear and quadratic terms, which involves finding the roots of the denominator.

6.1.4 Exercises

Exercise 6.1.5: Find the Laplace transform of $3 + t^5 + \sin(\pi t)$.

Exercise 6.1.6: Find the Laplace transform of $a + bt + ct^2$ for some constants a , b , and c .

Exercise 6.1.7:* Find the Laplace transform of $4(t+1)^2$.

Exercise 6.1.8: Find the Laplace transform of $A \cos(\omega t) + B \sin(\omega t)$.

Exercise 6.1.9: Find the Laplace transform of $\cos^2(\omega t)$.

Exercise 6.1.10: Find the inverse Laplace transform of $\frac{4}{s^2-9}$.

Exercise 6.1.11: Find the inverse Laplace transform of $\frac{2s}{s^2-1}$.

Exercise 6.1.12:* Find the inverse Laplace transform of $\frac{8}{s^3(s+2)}$.

Exercise 6.1.13: Find the inverse Laplace transform of $\frac{1}{(s-1)^2(s+1)}$.

Exercise 6.1.14: Find the Laplace transform of $f(t) = \begin{cases} t & \text{if } t \geq 1, \\ 0 & \text{if } t < 1. \end{cases}$

Exercise 6.1.15: Find the inverse Laplace transform of $\frac{s}{(s^2+s+2)(s+4)}$.

Exercise 6.1.16: Find the Laplace transform of $\sin(\omega(t-a))$.

Exercise 6.1.17:* Find the Laplace transform of te^{-t} (Hint: integrate by parts).

Exercise 6.1.18: Find the Laplace transform of $t \sin(\omega t)$. Hint: Several integrations by parts.

Exercise 6.1.19:* Find the Laplace transform of $\sin(t)e^{-t}$ (Hint: integrate by parts).

6.2 Transforms of derivatives and ODEs

Attribution: [JL], §6.2.

Learning Objectives

After this section, you will be able to:

- Relate the Laplace transform of a derivative to the transform of the original function,
- Use the Laplace transform to solve ordinary differential equations, and
- Find the Laplace transform of an integral using the properties of the transform.

6.2.1 Transforms of derivatives

Let us see how the Laplace transform is used for differential equations. First let us try to find the Laplace transform of a function that is a derivative. Suppose $g(t)$ is a differentiable function of exponential order, that is, $|g(t)| \leq M e^{ct}$ for some M and c . So $\mathcal{L}\{g(t)\}$ exists, and what is more, $\lim_{t \rightarrow \infty} e^{-st} g(t) = 0$ when $s > c$. Then

$$\mathcal{L}\{g'(t)\} = \int_0^\infty e^{-st} g'(t) dt = \left[e^{-st} g(t) \right]_{t=0}^\infty - \int_0^\infty (-s) e^{-st} g(t) dt = -g(0) + s\mathcal{L}\{g(t)\}.$$

We repeat this procedure for higher derivatives. The results are listed in [Table 6.2](#). The procedure also works for piecewise smooth functions, that is functions that are piecewise continuous with a piecewise continuous derivative.

$f(t)$	$\mathcal{L}\{f(t)\} = F(s)$
$g'(t)$	$sG(s) - g(0)$
$g''(t)$	$s^2G(s) - sg(0) - g'(0)$
$g'''(t)$	$s^3G(s) - s^2g(0) - sg'(0) - g''(0)$

Table 6.2: Laplace transforms of derivatives ($G(s) = \mathcal{L}\{g(t)\}$ as usual).

Exercise 6.2.1: Verify [Table 6.2](#).

6.2.2 Solving ODEs with the Laplace transform

Notice that the Laplace transform turns differentiation into multiplication by s . Let us see how to apply this fact to differential equations.

Example 6.2.1: Solve the differential equation

$$x''(t) + x(t) = \cos(2t), \quad x(0) = 0, \quad x'(0) = 1.$$

using the Laplace transform.

Solution: We will take the Laplace transform of both sides. By $X(s)$ we will, as usual, denote the Laplace transform of $x(t)$.

$$\begin{aligned} \mathcal{L}\{x''(t) + x(t)\} &= \mathcal{L}\{\cos(2t)\}, \\ s^2X(s) - sx(0) - x'(0) + X(s) &= \frac{s}{s^2 + 4}. \end{aligned}$$

We plug in the initial conditions now—this makes the computations more streamlined—to obtain

$$s^2X(s) - 1 + X(s) = \frac{s}{s^2 + 4}.$$

We solve for $X(s)$,

$$X(s) = \frac{s}{(s^2 + 1)(s^2 + 4)} + \frac{1}{s^2 + 1}.$$

We use partial fractions (exercise) to write

$$X(s) = \frac{1}{3} \frac{s}{s^2 + 1} - \frac{1}{3} \frac{s}{s^2 + 4} + \frac{1}{s^2 + 1}.$$

Now take the inverse Laplace transform to obtain

$$x(t) = \frac{1}{3} \cos(t) - \frac{1}{3} \cos(2t) + \sin(t).$$

\boxed{}

The procedure for linear constant coefficient equations is as follows. We take an ordinary differential equation in the time variable t . We apply the Laplace transform to transform the equation into an algebraic (non differential) equation in the frequency domain. All the $x(t)$, $x'(t)$, $x''(t)$, and so on, will be converted to $X(s)$, $sX(s) - x(0)$, $s^2X(s) - sx(0) - x'(0)$, and so on. We solve the equation for $X(s)$. Then taking the inverse transform, if possible, we find $x(t)$.

It should be noted that since not every function has a Laplace transform, not every equation can be solved in this manner. Also if the equation is not a linear constant coefficient ODE, then by applying the Laplace transform we may not obtain an algebraic equation.

6.2.3 Using the Heaviside function

Before we move on to more general equations than those we could solve before, we want to consider the Heaviside function. See [Figure 6.1](#) on the following page for the graph.

$$u(t) = \begin{cases} 0 & \text{if } t < 0, \\ 1 & \text{if } t \geq 0. \end{cases}$$

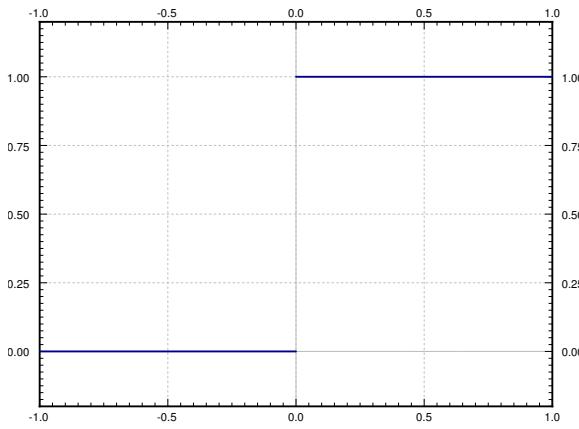


Figure 6.1: Plot of the Heaviside (unit step) function $u(t)$.

This function is useful for putting together functions, or cutting functions off. Most commonly it is used as $u(t - a)$ for some constant a . This just shifts the graph to the right by a . That is, it is a function that is 0 when $t < a$ and 1 when $t \geq a$. Suppose for example that $f(t)$ is a “signal” and you started receiving the signal $\sin t$ at time $t = \pi$. The function $f(t)$ should then be defined as

$$f(t) = \begin{cases} 0 & \text{if } t < \pi, \\ \sin t & \text{if } t \geq \pi. \end{cases}$$

Using the Heaviside function, $f(t)$ can be written as

$$f(t) = u(t - \pi) \sin t.$$

Similarly the step function that is 1 on the interval $[1, 2]$ and zero everywhere else can be written as

$$u(t - 1) - u(t - 2).$$

The Heaviside function is useful to define functions defined piecewise. If you want to define $f(t)$ such that $f(t) = t$ when t is in $[0, 1]$, $f(t) = -t + 2$ when t is in $[1, 2]$, and $f(t) = 0$ otherwise, then you can use the expression

$$f(t) = t(u(t) - u(t - 1)) + (-t + 2)(u(t - 1) - u(t - 2)).$$

Hence it is useful to know how the Heaviside function interacts with the Laplace transform. We have already seen that

$$\mathcal{L}\{u(t - a)\} = \frac{e^{-as}}{s}.$$

This can be generalized into a *shifting property* or *second shifting property*.

$$\mathcal{L}\{f(t - a)u(t - a)\} = e^{-as}\mathcal{L}\{f(t)\}.$$

(6.1)

Example 6.2.2: Suppose that the forcing function is not periodic. For example, suppose that we had a mass-spring system

$$x''(t) + x(t) = f(t), \quad x(0) = 0, \quad x'(0) = 0,$$

where $f(t) = 1$ if $1 \leq t < 5$ and zero otherwise. We could imagine a mass-spring system, where a rocket is fired for 4 seconds starting at $t = 1$. Or perhaps an RLC circuit, where the voltage is raised at a constant rate for 4 seconds starting at $t = 1$, and then held steady again starting at $t = 5$. Solve this differential equation using the Laplace transform.

Solution: We can write $f(t) = u(t - 1) - u(t - 5)$. We transform the equation and we plug in the initial conditions as before to obtain

$$s^2 X(s) + X(s) = \frac{e^{-s}}{s} - \frac{e^{-5s}}{s}.$$

We solve for $X(s)$ to obtain

$$X(s) = \frac{e^{-s}}{s(s^2 + 1)} - \frac{e^{-5s}}{s(s^2 + 1)}.$$

We leave it as an exercise to the reader to show that

$$\mathcal{L}^{-1} \left\{ \frac{1}{s(s^2 + 1)} \right\} = 1 - \cos t.$$

In other words $\mathcal{L}\{1 - \cos t\} = \frac{1}{s(s^2 + 1)}$. So using (6.1) we find

$$\mathcal{L}^{-1} \left\{ \frac{e^{-s}}{s(s^2 + 1)} \right\} = \mathcal{L}^{-1} \{ e^{-s} \mathcal{L}\{1 - \cos t\} \} = (1 - \cos(t - 1)) u(t - 1).$$

Similarly

$$\mathcal{L}^{-1} \left\{ \frac{e^{-5s}}{s(s^2 + 1)} \right\} = \mathcal{L}^{-1} \{ e^{-5s} \mathcal{L}\{1 - \cos t\} \} = (1 - \cos(t - 5)) u(t - 5).$$

Hence, the solution is

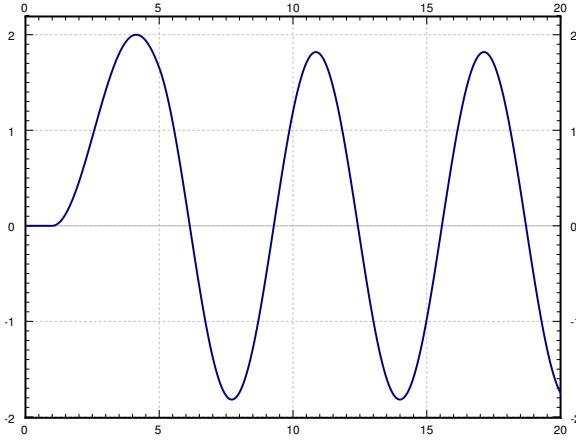
$$x(t) = (1 - \cos(t - 1)) u(t - 1) - (1 - \cos(t - 5)) u(t - 5).$$

The plot of this solution is given in Figure 6.2 on the next page. □

6.2.4 Transfer functions

The Laplace transform leads to the following useful concept for studying the steady state behavior of a linear system. Consider an equation of the form

$$Lx = f(t),$$

Figure 6.2: Plot of $x(t)$.

where L is a linear constant coefficient differential operator. Then $f(t)$ is usually thought of as input of the system and $x(t)$ is thought of as the output of the system. For example, for a mass-spring system the input is the forcing function and the output is the behavior of the mass. We would like to have a convenient way to study the behavior of the system for different inputs.

Let us suppose that all the initial conditions are zero and take the Laplace transform of the equation, we obtain the equation

$$A(s)X(s) = F(s).$$

Solving for the ratio $X(s)/F(s)$ we obtain the so-called *transfer function* $H(s) = 1/A(s)$, that is,

$$H(s) = \frac{X(s)}{F(s)}.$$

In other words, $X(s) = H(s)F(s)$. We obtain an algebraic dependence of the output of the system based on the input. We can now easily study the steady state behavior of the system given different inputs by simply multiplying by the transfer function.

Example 6.2.3: Find the transfer function for the differential equation $x'' + \omega_0^2 x = f(t)$ (assuming the initial conditions are zero).

Solution: First, we take the Laplace transform of the equation.

$$s^2 X(s) + \omega_0^2 X(s) = F(s).$$

Now we solve for the transfer function $X(s)/F(s)$.

$$H(s) = \frac{X(s)}{F(s)} = \frac{1}{s^2 + \omega_0^2}.$$

Let us see how to use the transfer function. Suppose we have the constant input $f(t) = 1$. Hence $F(s) = 1/s$, and

$$X(s) = H(s)F(s) = \frac{1}{s^2 + \omega_0^2} \frac{1}{s}.$$

Taking the inverse Laplace transform of $X(s)$ we obtain

$$x(t) = \frac{1 - \cos(\omega_0 t)}{\omega_0^2}.$$

□

6.2.5 Transforms of integrals

A feature of Laplace transforms is that it is also able to easily deal with integral equations. That is, equations in which integrals rather than derivatives of functions appear. The basic property, which can be proved by applying the definition and doing integration by parts, is

$$\mathcal{L} \left\{ \int_0^t f(\tau) d\tau \right\} = \frac{1}{s} F(s).$$

It is sometimes useful (e.g. for computing the inverse transform) to write this as

$$\int_0^t f(\tau) d\tau = \mathcal{L}^{-1} \left\{ \frac{1}{s} F(s) \right\}.$$

Example 6.2.4: To compute $\mathcal{L}^{-1} \left\{ \frac{1}{s(s^2+1)} \right\}$ we could proceed by applying this integration rule.

$$\mathcal{L}^{-1} \left\{ \frac{1}{s} \frac{1}{s^2+1} \right\} = \int_0^t \mathcal{L}^{-1} \left\{ \frac{1}{s^2+1} \right\} d\tau = \int_0^t \sin \tau d\tau = 1 - \cos t.$$

Example 6.2.5: An equation containing an integral of the unknown function is called an *integral equation*. For example, take

$$t^2 = \int_0^t e^\tau x(\tau) d\tau,$$

where we wish to solve for $x(t)$. We apply the Laplace transform and the shifting property to get

$$\frac{2}{s^3} = \frac{1}{s} \mathcal{L}\{e^t x(t)\} = \frac{1}{s} X(s-1),$$

where $X(s) = \mathcal{L}\{x(t)\}$. Thus

$$X(s-1) = \frac{2}{s^2} \quad \text{or} \quad X(s) = \frac{2}{(s+1)^2}.$$

We use the shifting property again

$$x(t) = 2e^{-t}t.$$

6.2.6 Exercises

Exercise 6.2.2: Using the Heaviside function write down the piecewise function that is 0 for $t < 0$, t^2 for t in $[0, 1]$ and t for $t > 1$.

Exercise 6.2.3:* Using the Heaviside function $u(t)$, write down the function

$$f(t) = \begin{cases} 0 & \text{if } t < 1, \\ t - 1 & \text{if } 1 \leq t < 2, \\ 1 & \text{if } 2 \leq t. \end{cases}$$

Exercise 6.2.4: Using the Laplace transform solve

$$mx'' + cx' + kx = 0, \quad x(0) = a, \quad x'(0) = b,$$

where $m > 0$, $c > 0$, $k > 0$, and $c^2 - 4km > 0$ (system is overdamped).

Exercise 6.2.5: Using the Laplace transform solve

$$mx'' + cx' + kx = 0, \quad x(0) = a, \quad x'(0) = b,$$

where $m > 0$, $c > 0$, $k > 0$, and $c^2 - 4km < 0$ (system is underdamped).

Exercise 6.2.6: Using the Laplace transform solve

$$mx'' + cx' + kx = 0, \quad x(0) = a, \quad x'(0) = b,$$

where $m > 0$, $c > 0$, $k > 0$, and $c^2 = 4km$ (system is critically damped).

Exercise 6.2.7: Solve $x'' + x = u(t - 1)$ for initial conditions $x(0) = 0$ and $x'(0) = 0$.

Exercise 6.2.8:* Solve $x'' - x = (t^2 - 1)u(t - 1)$ for initial conditions $x(0) = 1$, $x'(0) = 2$ using the Laplace transform.

Exercise 6.2.9: Show the differentiation of the transform property. Suppose $\mathcal{L}\{f(t)\} = F(s)$, then show

$$\mathcal{L}\{-tf(t)\} = F'(s).$$

Hint: Differentiate under the integral sign.

Exercise 6.2.10: Solve $x''' + x = t^3u(t - 1)$ for initial conditions $x(0) = 1$ and $x'(0) = 0$, $x''(0) = 0$.

Exercise 6.2.11: Show the second shifting property: $\mathcal{L}\{f(t - a)u(t - a)\} = e^{-as}\mathcal{L}\{f(t)\}$.

Exercise 6.2.12: Let us think of the mass-spring system with a rocket from [Example 6.2.2](#). We noticed that the solution kept oscillating after the rocket stopped running. The amplitude of the oscillation depends on the time that the rocket was fired (for 4 seconds in the example).

- a) Find a formula for the amplitude of the resulting oscillation in terms of the amount of time the rocket is fired.
- b) Is there a nonzero time (if so what is it?) for which the rocket fires and the resulting oscillation has amplitude 0 (the mass is not moving)?

Exercise 6.2.13: Define

$$f(t) = \begin{cases} (t-1)^2 & \text{if } 1 \leq t < 2, \\ 3-t & \text{if } 2 \leq t < 3, \\ 0 & \text{otherwise.} \end{cases}$$

- a) Sketch the graph of $f(t)$.
- b) Write down $f(t)$ using the Heaviside function.
- c) Solve $x'' + x = f(t)$, $x(0) = 0$, $x'(0) = 0$ using Laplace transform.

Exercise 6.2.14: Find the transfer function for $mx'' + cx' + kx = f(t)$ (assuming the initial conditions are zero).

Exercise 6.2.15:* Find the transfer function for $x' + x = f(t)$ (assuming the initial conditions are zero).

6.3 Convolution

Attribution: [JL], §6.3.

Learning Objectives

After this section, you will be able to:

- Compute the convolution of two functions,
- Relate the convolution of two functions to the Laplace transform of those functions, and
- Use the Laplace transform to solve more complicated differential equations involving products.

6.3.1 The convolution

The Laplace transformation of a product is not the product of the transforms. All hope is not lost however. We simply have to use a different type of a “product.” Take two functions $f(t)$ and $g(t)$ defined for $t \geq 0$, and define the *convolution** of $f(t)$ and $g(t)$ as

$$(f * g)(t) \stackrel{\text{def}}{=} \int_0^t f(\tau)g(t - \tau) d\tau. \quad (6.2)$$

As you can see, the convolution of two functions of t is another function of t .

Example 6.3.1: Take $f(t) = e^t$ and $g(t) = t$ for $t \geq 0$. Then

$$(f * g)(t) = \int_0^t e^\tau(t - \tau) d\tau = e^t - t - 1.$$

To solve the integral we did one integration by parts.

Example 6.3.2: Take $f(t) = \sin(\omega t)$ and $g(t) = \cos(\omega t)$ for $t \geq 0$. Then

$$(f * g)(t) = \int_0^t \sin(\omega\tau) \cos(\omega(t - \tau)) d\tau.$$

Apply the identity

$$\cos(\theta)\sin(\psi) = \frac{1}{2} (\sin(\theta + \psi) - \sin(\theta - \psi)),$$

*For those that have seen convolution before, you may have seen it defined as $(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau) d\tau$. This definition agrees with (6.2) if you define $f(t)$ and $g(t)$ to be zero for $t < 0$. When discussing the Laplace transform the definition we gave is sufficient. Convolution does occur in many other applications, however, where you may have to use the more general definition with infinities.

to get

$$\begin{aligned}(f * g)(t) &= \int_0^t \frac{1}{2} (\sin(\omega t) - \sin(\omega t - 2\omega\tau)) d\tau \\&= \left[\frac{1}{2} \tau \sin(\omega t) + \frac{1}{4\omega} \cos(2\omega\tau - \omega t) \right]_{\tau=0}^t \\&= \frac{1}{2} t \sin(\omega t).\end{aligned}$$

The formula holds only for $t \geq 0$. The functions f , g , and $f * g$ are undefined for $t < 0$.

Convolution has many properties that make it behave like a product. Let c be a constant and f , g , and h be functions. Then

$$\begin{aligned}f * g &= g * f, \\(cf) * g &= f * (cg) = c(f * g), \\(f * g) * h &= f * (g * h).\end{aligned}$$

The most interesting property for us is the following theorem.

Theorem 6.3.1

Let $f(t)$ and $g(t)$ be of exponential order, then

$$\mathcal{L}\{(f * g)(t)\} = \mathcal{L}\left\{\int_0^t f(\tau)g(t-\tau) d\tau\right\} = \mathcal{L}\{f(t)\}\mathcal{L}\{g(t)\}.$$

In other words, the Laplace transform of a convolution is the product of the Laplace transforms. The simplest way to use this result is in reverse.

Example 6.3.3: Suppose we have the function of s defined by

$$\frac{1}{(s+1)s^2} = \frac{1}{s+1} \frac{1}{s^2}.$$

We recognize the two entries of [Table 6.2](#). That is,

$$\mathcal{L}^{-1}\left\{\frac{1}{s+1}\right\} = e^{-t} \quad \text{and} \quad \mathcal{L}^{-1}\left\{\frac{1}{s^2}\right\} = t.$$

Therefore,

$$\mathcal{L}^{-1}\left\{\frac{1}{s+1} \frac{1}{s^2}\right\} = \int_0^t \tau e^{-(t-\tau)} d\tau = e^{-t} + t - 1.$$

The calculation of the integral involved an integration by parts.

6.3.2 Solving ODEs

The next example demonstrates the full power of the convolution and the Laplace transform. We can give the solution to the forced oscillation problem for any forcing function as a definite integral.

Example 6.3.4: Find the solution to

$$x'' + \omega_0^2 x = f(t), \quad x(0) = 0, \quad x'(0) = 0,$$

for an arbitrary function $f(t)$.

Solution: We first apply the Laplace transform to the equation. Denote the transform of $x(t)$ by $X(s)$ and the transform of $f(t)$ by $F(s)$ as usual.

$$s^2 X(s) + \omega_0^2 X(s) = F(s),$$

or in other words

$$X(s) = F(s) \frac{1}{s^2 + \omega_0^2}.$$

We know

$$\mathcal{L}^{-1} \left\{ \frac{1}{s^2 + \omega_0^2} \right\} = \frac{\sin(\omega_0 t)}{\omega_0}.$$

Therefore,

$$x(t) = \int_0^t f(\tau) \frac{\sin(\omega_0(t-\tau))}{\omega_0} d\tau,$$

or if we reverse the order

$$x(t) = \int_0^t \frac{\sin(\omega_0 \tau)}{\omega_0} f(t-\tau) d\tau.$$

□

Notice one more feature of this example. We can now see how Laplace transform handles resonance. Suppose that $f(t) = \cos(\omega_0 t)$. Then

$$x(t) = \int_0^t \frac{\sin(\omega_0 \tau)}{\omega_0} \cos(\omega_0(t-\tau)) d\tau = \frac{1}{\omega_0} \int_0^t \sin(\omega_0 \tau) \cos(\omega_0(t-\tau)) d\tau.$$

We have computed the convolution of sine and cosine in [Example 6.3.2](#). Hence

$$x(t) = \left(\frac{1}{\omega_0} \right) \left(\frac{1}{2} t \sin(\omega_0 t) \right) = \frac{1}{2\omega_0} t \sin(\omega_0 t).$$

Note the t in front of the sine. The solution, therefore, grows without bound as t gets large, meaning we get resonance.

Similarly, we can solve any constant coefficient equation with an arbitrary forcing function $f(t)$ as a definite integral using convolution. A definite integral, rather than a closed form solution, is usually enough for most practical purposes. It is not hard to numerically evaluate a definite integral.

6.3.3 Volterra integral equation

A common integral equation is the *Volterra integral equation*^{*}

$$x(t) = f(t) + \int_0^t g(t-\tau)x(\tau) d\tau,$$

where $f(t)$ and $g(t)$ are known functions and $x(t)$ is an unknown we wish to solve for. To find $x(t)$, we apply the Laplace transform to the equation to obtain

$$X(s) = F(s) + G(s)X(s),$$

where $X(s)$, $F(s)$, and $G(s)$ are the Laplace transforms of $x(t)$, $f(t)$, and $g(t)$ respectively. We find

$$X(s) = \frac{F(s)}{1 - G(s)}.$$

To find $x(t)$ we now need to find the inverse Laplace transform of $X(s)$.

Example 6.3.5: Solve

$$x(t) = e^{-t} + \int_0^t \sinh(t-\tau)x(\tau) d\tau.$$

Solution: We apply Laplace transform to obtain

$$X(s) = \frac{1}{s+1} + \frac{1}{s^2-1}X(s),$$

or

$$X(s) = \frac{\frac{1}{s+1}}{1 - \frac{1}{s^2-1}} = \frac{s-1}{s^2-2} = \frac{s}{s^2-2} - \frac{1}{s^2-2}.$$

It is not hard to apply Table 6.1 on page 359 to find

$$x(t) = \cosh(\sqrt{2}t) - \frac{1}{\sqrt{2}}\sinh(\sqrt{2}t).$$

□

6.3.4 Exercises

Exercise 6.3.1: Let $f(t) = t^2$ for $t \geq 0$, and $g(t) = u(t-1)$. Compute $f * g$.

Exercise 6.3.2: Let $f(t) = t$ for $t \geq 0$, and $g(t) = \sin t$ for $t \geq 0$. Compute $f * g$.

Exercise 6.3.3:* Let $f(t) = \cos t$ for $t \geq 0$, and $g(t) = e^{-t}$. Compute $f * g$.

Exercise 6.3.4: Find the solution to

$$mx'' + cx' + kx = f(t), \quad x(0) = 0, \quad x'(0) = 0,$$

for an arbitrary function $f(t)$, where $m > 0$, $c > 0$, $k > 0$, and $c^2 - 4km > 0$ (system is overdamped). Write the solution as a definite integral.

*Named for the Italian mathematician Vito Volterra (1860–1940).

Exercise 6.3.5: Find the solution to

$$mx'' + cx' + kx = f(t), \quad x(0) = 0, \quad x'(0) = 0,$$

for an arbitrary function $f(t)$, where $m > 0$, $c > 0$, $k > 0$, and $c^2 - 4km < 0$ (system is underdamped). Write the solution as a definite integral.

Exercise 6.3.6: Find the solution to

$$mx'' + cx' + kx = f(t), \quad x(0) = 0, \quad x'(0) = 0,$$

for an arbitrary function $f(t)$, where $m > 0$, $c > 0$, $k > 0$, and $c^2 = 4km$ (system is critically damped). Write the solution as a definite integral.

Exercise 6.3.7: Solve

$$x(t) = e^{-t} + \int_0^t \cos(t - \tau)x(\tau) d\tau.$$

Exercise 6.3.8:* Solve $x'' + x = \sin t$, $x(0) = 0$, $x'(0) = 0$ using convolution.

Exercise 6.3.9: Solve

$$x(t) = \cos t + \int_0^t \cos(t - \tau)x(\tau) d\tau.$$

Exercise 6.3.10:* Solve $x''' + x' = f(t)$, $x(0) = 0$, $x'(0) = 0$, $x''(0) = 0$ using convolution. Write the result as a definite integral.

Exercise 6.3.11: Compute $\mathcal{L}^{-1}\left\{\frac{s}{(s^2+4)^2}\right\}$ using convolution.

Exercise 6.3.12:* Compute $\mathcal{L}^{-1}\left\{\frac{5}{s^4+s^2}\right\}$ using convolution.

Exercise 6.3.13: Write down the solution to $x'' - 2x = e^{-t^2}$, $x(0) = 0$, $x'(0) = 0$ as a definite integral. Hint: Do not try to compute the Laplace transform of e^{-t^2} .

6.4 Dirac delta and impulse response

Attribution: [JL], §6.4.

Learning Objectives

After this section, you will be able to:

- Understand the Dirac delta function as an impulse test for a differential equation, and
- Use the Laplace transform to solve differential equations using the Dirac delta function.

6.4.1 Rectangular pulse

Often in applications we study a physical system by putting in a short pulse and then seeing what the system does. The resulting behavior is often called *impulse response*. Let us see what we mean by a pulse. The simplest kind of a pulse is a simple rectangular pulse defined by

$$\varphi(t) = \begin{cases} 0 & \text{if } t < a, \\ M & \text{if } a \leq t < b, \\ 0 & \text{if } b \leq t. \end{cases}$$

See Figure 6.3 for a graph.

Notice that

$$\varphi(t) = M(u(t - a) - u(t - b)),$$

where $u(t)$ is the unit step function.

Let us take the Laplace transform of a square pulse,

$$\begin{aligned} \mathcal{L}\{\varphi(t)\} &= \mathcal{L}\{M(u(t - a) - u(t - b))\} \\ &= M \frac{e^{-as} - e^{-bs}}{s}. \end{aligned}$$

For simplicity we let $a = 0$, and it is convenient to set $M = 1/b$ to have

$$\int_0^\infty \varphi(t) dt = 1.$$

That is, to have the pulse have “unit mass.” For such a pulse we compute

$$\mathcal{L}\{\varphi(t)\} = \mathcal{L}\left\{\frac{u(t) - u(t - b)}{b}\right\} = \frac{1 - e^{-bs}}{bs}.$$

We generally want b to be very small. That is, we wish to have the pulse be very short and very tall. By letting b go to zero we arrive at the concept of the Dirac delta function.

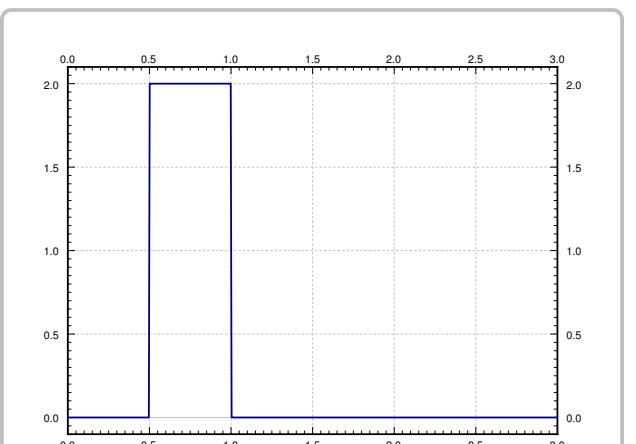


Figure 6.3: Sample square pulse with $a = 0.5$, $b = 1$ and $M = 2$.

6.4.2 The delta function

The *Dirac delta function** is not exactly a function; it is sometimes called a *generalized function*. We avoid unnecessary details and simply say that it is an object that does not really make sense unless we integrate it. The motivation is that we would like a “function” $\delta(t)$ such that for any continuous function $f(t)$ we have

$$\int_{-\infty}^{\infty} \delta(t)f(t) dt = f(0).$$

The formula should hold if we integrate over any interval that contains 0, not just $(-\infty, \infty)$. So $\delta(t)$ is a “function” with all its “mass” at the single point $t = 0$. In other words, for any interval $[c, d]$

$$\int_c^d \delta(t) dt = \begin{cases} 1 & \text{if the interval } [c, d] \text{ contains 0, i.e. } c \leq 0 \leq d, \\ 0 & \text{otherwise.} \end{cases}$$

Unfortunately there is no such function in the classical sense. You could informally think that $\delta(t)$ is zero for $t \neq 0$ and somehow infinite at $t = 0$.

A good way to think about $\delta(t)$ is as a limit of short pulses whose integral is 1. For example, suppose that we have a square pulse $\varphi(t)$ as above with $a = 0$, $M = 1/b$, that is $\varphi(t) = \frac{u(t)-u(t-b)}{b}$. Compute

$$\int_{-\infty}^{\infty} \varphi(t)f(t) dt = \int_{-\infty}^{\infty} \frac{u(t) - u(t-b)}{b} f(t) dt = \frac{1}{b} \int_0^b f(t) dt.$$

If $f(t)$ is continuous at $t = 0$, then for very small b , the function $f(t)$ is approximately equal to $f(0)$ on the interval $[0, b]$. We approximate the integral

$$\frac{1}{b} \int_0^b f(t) dt \approx \frac{1}{b} \int_0^b f(0) dt = f(0).$$

Hence,

$$\lim_{b \rightarrow 0} \int_{-\infty}^{\infty} \varphi(t)f(t) dt = \lim_{b \rightarrow 0} \frac{1}{b} \int_0^b f(t) dt = f(0).$$

Let us therefore accept $\delta(t)$ as an object that is possible to integrate. We often want to shift δ to another point, for example $\delta(t-a)$. In that case we have

$$\int_{-\infty}^{\infty} \delta(t-a)f(t) dt = f(a).$$

Note that $\delta(a-t)$ is the same object as $\delta(t-a)$. In other words, the convolution of $\delta(t)$ with $f(t)$ is again $f(t)$,

$$(f * \delta)(t) = \int_0^t \delta(t-s)f(s) ds = f(t).$$

*Named after the English physicist and mathematician Paul Adrien Maurice Dirac (1902–1984).

As we can integrate $\delta(t)$, let us compute its Laplace transform.

$$\mathcal{L}\{\delta(t-a)\} = \int_0^\infty e^{-st}\delta(t-a) dt = e^{-as}.$$

In particular,

$$\mathcal{L}\{\delta(t)\} = 1.$$

Remark 6.4.1: Notice that the Laplace transform of $\delta(t-a)$ looks like the Laplace transform of the derivative of the Heaviside function $u(t-a)$, if we could differentiate the Heaviside function. First notice

$$\mathcal{L}\{u(t-a)\} = \frac{e^{-as}}{s}.$$

To obtain what the Laplace transform of the derivative would be we multiply by s , to obtain e^{-as} , which is the Laplace transform of $\delta(t-a)$. We see the same thing using integration,

$$\int_0^t \delta(s-a) ds = u(t-a).$$

So in a certain sense

$$\text{“} \frac{d}{dt} [u(t-a)] = \delta(t-a). \text{”}$$

This line of reasoning allows us to talk about derivatives of functions with jump discontinuities. We can think of the derivative of the Heaviside function $u(t-a)$ as being somehow infinite at a , which is precisely our intuitive understanding of the delta function.

Example 6.4.1: Let us compute $\mathcal{L}^{-1}\left\{\frac{s+1}{s}\right\}$. So far we have always looked at proper rational functions in the s variable. That is, the numerator was always of lower degree than the denominator. Not so with $\frac{s+1}{s}$. We write,

$$\mathcal{L}^{-1}\left\{\frac{s+1}{s}\right\} = \mathcal{L}^{-1}\left\{1 + \frac{1}{s}\right\} = \mathcal{L}^{-1}\{1\} + \mathcal{L}^{-1}\left\{\frac{1}{s}\right\} = \delta(t) + 1.$$

The resulting object is a generalized function and only makes sense when put underneath an integral.

6.4.3 Impulse response

As we said before, in the differential equation $Lx = f(t)$, we think of $f(t)$ as input, and $x(t)$ as the output. Often it is important to find the response to an impulse, and then we use the delta function in place of $f(t)$. The solution to

$$Lx = \delta(t)$$

is called the *impulse response*.

Example 6.4.2: Solve (find the impulse response)

$$x'' + \omega_0^2 x = \delta(t), \quad x(0) = 0, \quad x'(0) = 0. \quad (6.3)$$

Solution: We first apply the Laplace transform to the equation. Denote the transform of $x(t)$ by $X(s)$.

$$s^2 X(s) + \omega_0^2 X(s) = 1, \quad \text{and so} \quad X(s) = \frac{1}{s^2 + \omega_0^2}.$$

Taking the inverse Laplace transform we obtain

$$x(t) = \frac{\sin(\omega_0 t)}{\omega_0}.$$

]

Let us notice something about the example above. We showed before that when the input is $f(t)$, then the solution to $Lx = f(t)$ is given by

$$x(t) = \int_0^t f(\tau) \frac{\sin(\omega_0(t-\tau))}{\omega_0} d\tau.$$

That is, the solution for an arbitrary input is given as convolution with the impulse response. Let us see why. The key is to notice that for functions $x(t)$ and $f(t)$,

$$(x * f)''(t) = \frac{d^2}{dt^2} \left[\int_0^t f(\tau)x(t-\tau) d\tau \right] = \int_0^t f(\tau)x''(t-\tau) d\tau = (x'' * f)(t).$$

We simply differentiate twice under the integral*, the details are left as an exercise. If we convolve the entire equation (6.3), the left-hand side becomes

$$(x'' + \omega_0^2 x) * f = (x'' * f) + \omega_0^2(x * f) = (x * f)'' + \omega_0^2(x * f).$$

The right-hand side becomes

$$(\delta * f)(t) = f(t).$$

Therefore $y(t) = (x * f)(t)$ is the solution to

$$y'' + \omega_0^2 y = f(t).$$

This procedure works in general for other linear equations $Lx = f(t)$. If you determine the impulse response, you also know how to obtain the output $x(t)$ for any input $f(t)$ by simply convolving the impulse response and the input $f(t)$.

6.4.4 Three-point beam bending

Let us give another quite different example where delta functions turn up. In this case representing point loads on a steel beam. Suppose we have a beam of length L , resting on two simple supports at the ends. Let x denote the position on the beam, and let $y(x)$

*You should really think of the integral going over $(-\infty, \infty)$ rather than over $[0, t]$ and simply assume that $f(t)$ and $x(t)$ are continuous and zero for negative t .

denote the deflection of the beam in the vertical direction. The deflection $y(x)$ satisfies the *Euler–Bernoulli equation*^{*},

$$EI \frac{d^4y}{dx^4} = F(x),$$

where E and I are constants[†] and $F(x)$ is the force applied per unit length at position x . The situation we are interested in is when the force is applied at a single point as in [Figure 6.4](#).

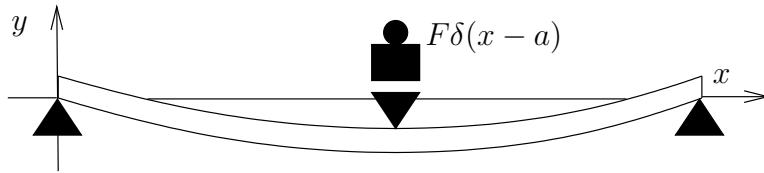


Figure 6.4: Three-point bending.

In this case the equation becomes

$$EI \frac{d^4y}{dx^4} = -F\delta(x - a),$$

where $x = a$ is the point where the mass is applied. F is the force applied and the minus sign indicates that the force is downward, that is, in the negative y direction. The end points of the beam satisfy the conditions,

$$\begin{aligned} y(0) &= 0, & y''(0) &= 0, \\ y(L) &= 0, & y''(L) &= 0. \end{aligned}$$

Example 6.4.3: Suppose that length of the beam is 2, and suppose that $EI = 1$ for simplicity. Further suppose that the force $F = 1$ is applied at $x = 1$. That is, we have the equation

$$\frac{d^4y}{dx^4} = -\delta(x - 1),$$

and the endpoint conditions are

$$y(0) = 0, \quad y''(0) = 0, \quad y(2) = 0, \quad y''(2) = 0.$$

Solution: We could integrate, but using the Laplace transform is even easier. We apply the transform in the x variable rather than the t variable. Let us again denote the transform of $y(x)$ as $Y(s)$.

$$s^4 Y(s) - s^3 y(0) - s^2 y'(0) - s y''(0) - y'''(0) = -e^{-s}.$$

*Named for the Swiss mathematicians [Jacob Bernoulli](#) (1654–1705), [Daniel Bernoulli](#) (1700–1782), the nephew of Jacob, and [Leonhard Paul Euler](#) (1707–1783).

† E is the elastic modulus and I is the second moment of area. Let us not worry about the details and simply think of these as some given constants.

We notice that $y(0) = 0$ and $y''(0) = 0$. Let us call $C_1 = y'(0)$ and $C_2 = y'''(0)$. We solve for $Y(s)$,

$$Y(s) = \frac{-e^{-s}}{s^4} + \frac{C_1}{s^2} + \frac{C_2}{s^4}.$$

We take the inverse Laplace transform utilizing the second shifting property (6.1) to take the inverse of the first term.

$$y(x) = \frac{-(x-1)^3}{6}u(x-1) + C_1x + \frac{C_2}{6}x^3.$$

We still need to apply two of the endpoint conditions. As the conditions are at $x = 2$ we can simply replace $u(x-1) = 1$ when taking the derivatives. Therefore,

$$0 = y(2) = \frac{-(2-1)^3}{6} + C_1(2) + \frac{C_2}{6}2^3 = \frac{-1}{6} + 2C_1 + \frac{4}{3}C_2,$$

and

$$0 = y''(2) = \frac{-3 \cdot 2 \cdot (2-1)}{6} + \frac{C_2}{6}3 \cdot 2 \cdot 2 = -1 + 2C_2.$$

Hence $C_2 = \frac{1}{2}$ and solving for C_1 using the first equation we obtain $C_1 = \frac{-1}{4}$. Our solution for the beam deflection is

$$y(x) = \frac{-(x-1)^3}{6}u(x-1) - \frac{x}{4} + \frac{x^3}{12}.$$

\(\square\)

6.4.5 Exercises

Exercise 6.4.1: Solve (find the impulse response) $x'' + x' + x = \delta(t)$, $x(0) = 0$, $x'(0) = 0$.

Exercise 6.4.2: Solve (find the impulse response) $x'' + 2x' + x = \delta(t)$, $x(0) = 0$, $x'(0) = 0$.

Exercise 6.4.3:* Solve (find the impulse response) $x'' = \delta(t)$, $x(0) = 0$, $x'(0) = 0$.

Exercise 6.4.4:* Solve (find the impulse response) $x' + ax = \delta(t)$, $x(0) = 0$, $x'(0) = 0$.

Exercise 6.4.5: A pulse can come later and can be bigger. Solve $x'' + 4x = 4\delta(t-1)$, $x(0) = 0$, $x'(0) = 0$.

Exercise 6.4.6: Suppose that $f(t)$ and $g(t)$ are differentiable functions and suppose that $f(t) = g(t) = 0$ for all $t \leq 0$. Show that

$$(f * g)'(t) = (f' * g)(t) = (f * g')(t).$$

Exercise 6.4.7: Suppose that $Lx = \delta(t)$, $x(0) = 0$, $x'(0) = 0$, has the solution $x = e^{-t}$ for $t > 0$. Find the solution to $Lx = t^2$, $x(0) = 0$, $x'(0) = 0$ for $t > 0$.

Exercise 6.4.8:* Suppose that $Lx = \delta(t)$, $x(0) = 0$, $x'(0) = 0$, has the solution $x(t) = \cos(t)$ for $t > 0$. Find (in closed form) the solution to $Lx = \sin(t)$, $x(0) = 0$, $x'(0) = 0$ for $t > 0$.

Exercise 6.4.9: Compute $\mathcal{L}^{-1} \left\{ \frac{s^2+s+1}{s^2} \right\}$.

Exercise 6.4.10:* Compute $\mathcal{L}^{-1} \left\{ \frac{s^2}{s^2+1} \right\}$.

Exercise 6.4.11:* Compute $\mathcal{L}^{-1} \left\{ \frac{3s^2e^{-s}+2}{s^2} \right\}$.

Exercise 6.4.12 (challenging): Solve *Example 6.4.3* via integrating 4 times in the x variable.

Exercise 6.4.13: Suppose we have a beam of length 1 simply supported at the ends and suppose that force $F = 1$ is applied at $x = \frac{3}{4}$ in the downward direction. Suppose that $EI = 1$ for simplicity. Find the beam deflection $y(x)$.

6.5 Solving PDEs with the Laplace transform

Attribution: [JL], §6.5.

Learning Objectives

After this section, you will be able to:

- Use the Laplace transform in one variable to solve PDE

The Laplace transform comes from the same family of transforms as does the Fourier series*, which we used in chapter 9 to solve partial differential equations (PDEs). It is therefore not surprising that we can also solve PDEs with the Laplace transform.

Given a PDE in two independent variables x and t , we use the Laplace transform on one of the variables (taking the transform of everything in sight), and derivatives in that variable become multiplications by the transformed variable s . The PDE becomes an ODE, which we solve. Afterwards we invert the transform to find a solution to the original problem. It is best to see the procedure on an example.

Example 6.5.1: Consider the first order PDE

$$y_t = -\alpha y_x, \quad \text{for } x > 0, \quad t > 0,$$

with side conditions

$$y(0, t) = C, \quad y(x, 0) = 0.$$

This equation is called the *convection equation* or sometimes the *transport equation*. See Figure 6.5 for a diagram of the setup.

A physical setup of this equation is a river of solid goo, as we do not want anything to diffuse. The function y is the concentration of some toxic substance†. The variable x denotes position where $x = 0$ is the location of a factory spewing the toxic substance into the river. The toxic substance flows into the river so that at $x = 0$ the concentration is always C . We wish to see what happens past the factory, that is at $x > 0$. Let t be the time, and assume the factory started operations at $t = 0$, so that at $t = 0$ the river is just pure goo.

Consider a function of two variables $y(x, t)$.

Let us fix x and transform the t variable. For convenience, we treat the transformed s variable

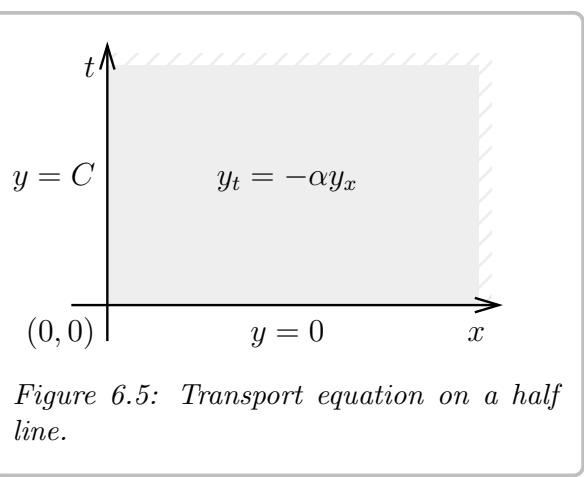


Figure 6.5: Transport equation on a half line.

*There is a corresponding Fourier transform on the real line as well that looks sort of like the Laplace transform.

†It's a river of goo already, we're not hurting the environment much more.

as a parameter, since there are no derivatives in s . That is, we write $Y(x)$ for the transformed function, and treat it as a function of x , leaving s as a parameter.

$$Y(x) = \mathcal{L}\{y(x, t)\} = \int_0^\infty y(x, t)e^{-st} ds.$$

The transform of a derivative with respect to x is just differentiating the transformed function:

$$\mathcal{L}\{y_x(x, t)\} = \int_0^\infty y_x(x, t)e^{-st} ds = \frac{d}{dx} \left[\int_0^\infty y(x, t)e^{-st} ds \right] = Y'(x).$$

To transform the derivative in t (the variable being transformed), we use the rules from § 6.2:

$$\mathcal{L}\{y_t(x, t)\} = sY(x) - y(x, 0).$$

In our specific case, $y(x, 0) = 0$, and so $\mathcal{L}\{y_t(x, t)\} = sY(x)$. We transform the equation to find

$$sY(x) = -\alpha Y'(x).$$

This ODE needs an initial condition. The initial condition is the other side condition of the PDE, the one that depends on x . Everything is transformed, so we must also transform this condition

$$Y(0) = \mathcal{L}\{y(0, t)\} = \mathcal{L}\{C\} = \frac{C}{s}.$$

We solve the ODE problem $sY(x) = -\alpha Y'(x)$, $Y(0) = \frac{C}{s}$, to find

$$Y(x) = \frac{C}{s} e^{-\frac{s}{\alpha}x}.$$

We are not done, we have $Y(x)$, but we really want $y(x, t)$. We transform the s variable back to t . Let

$$u(t) = \begin{cases} 0 & \text{if } t < 0, \\ 1 & \text{otherwise} \end{cases}$$

be the Heaviside function. As

$$\mathcal{L}\{u(t - a)\} = \int_0^\infty u(t - a) e^{-st} dt = \int_a^\infty e^{-st} dt = \frac{e^{-as}}{s},$$

then

$$y(x, t) = \mathcal{L}^{-1} \left\{ \frac{C}{s} e^{-\frac{s}{\alpha}x} \right\} = Cu(t - x/\alpha).$$

In other words,

$$y(x, t) = \begin{cases} 0 & \text{if } t < x/\alpha, \\ C & \text{otherwise.} \end{cases}$$

See Figure 6.6 on the next page for a diagram of this solution. The line of slope $1/\alpha$ indicates the wavefront of the toxic substance in the picture as it is leaving the factory. What the equation does is simply move the initial condition to the right at speed α .

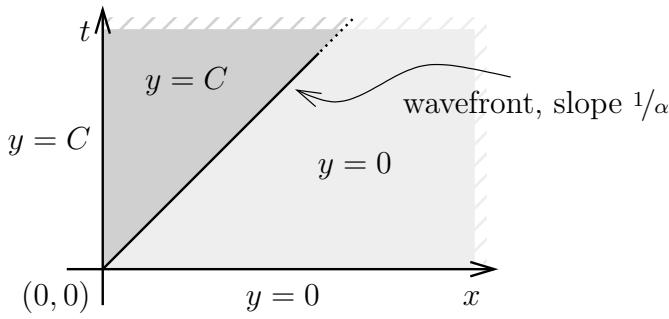


Figure 6.6: Wavefront of toxic substance is a line of slope $1/\alpha$.

Shhh... y is not differentiable, it is not even continuous (nobody ever seems to notice). How could we plug something that's not differentiable into the equation? Well, just think of a differentiable function very very close to y . Or, if you recognize the derivative of the Heaviside function as the delta function, then all is well too:

$$y_t(x, t) = \frac{\partial}{\partial t} [Cu(t - x/\alpha)] = Cu'(t - x/\alpha) = C\delta(t - x/\alpha)$$

and

$$y_x(x, t) = \frac{\partial}{\partial x} [Cu(t - x/\alpha)] = -\frac{C}{\alpha}u'(t - x/\alpha) = -\frac{C}{\alpha}\delta(t - x/\alpha).$$

So $y_t = -\alpha y_x$.

Laplace equation is very good with constant coefficient equations. One advantage of Laplace is that it easily handles nonhomogeneous side conditions. Let us try a more complicated example.

Example 6.5.2: Use the Laplace transform to solve

$$\begin{aligned} y_t + y_x + y &= 0, & \text{for } x > 0, t > 0, \\ y(0, t) &= \sin(t), & y(x, 0) = 0. \end{aligned}$$

Solution: Again, we transform t , and we write $Y(x)$ for the transformed function. As $y(x, 0) = 0$, we find

$$sY(x) + Y'(x) + Y(x) = 0, \quad Y(0) = \frac{1}{s^2 + 1}.$$

The solution of the transformed equation is

$$Y(x) = \frac{1}{s^2 + 1}e^{-(s+1)x} = \frac{1}{s^2 + 1}e^{-xs}e^{-x}.$$

Using the second shifting property (6.1) and linearity of the transform, we obtain the solution

$$y(x, t) = e^{-x} \sin(t - x)u(t - x).$$

□

We can also detect when the problem is *ill-posed* in the sense that it has no solution. Let us change the equation to

$$\begin{aligned} -y_t + y_x &= 0, & \text{for } x > 0, t > 0, \\ y(0, t) &= \sin(t), & y(x, 0) = 0. \end{aligned}$$

Then the problem has no solution. In the language of first order PDE, which will be discussed in § 10.1, this can be seen as follows. The characteristic curves are $t = -x + C$. If τ is the characteristic coordinate, then we find the equation $y_\tau = 0$ along the curve, meaning a solution is constant along characteristic curves. But these curves intersect both the x -axis and the t -axis. For example, the curve $t = -x + 1$ intersects at $(1, 0)$ and $(0, 1)$. The solution is constant along the curve so $y(1, 0)$ should equal $y(0, 1)$. But $y(1, 0) = 0$ and $y(0, 1) = \sin(1) \neq 0$. See Figure 6.7.

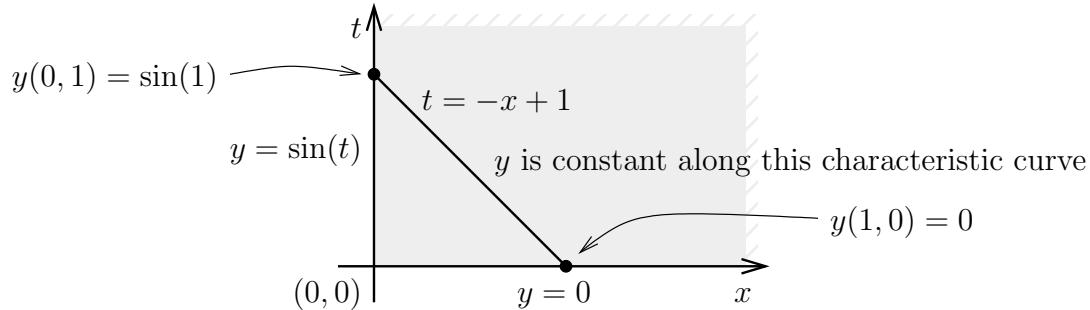


Figure 6.7: Ill-posed problem.

Now consider the transform. The transformed problem is

$$-sY(x) + Y'(x) = 0, \quad Y(0) = \frac{1}{s^2 + 1},$$

and the solution ought to be

$$Y(x) = \frac{1}{s^2 + 1} e^{sx}.$$

Importantly, this Laplace transform does not decay to zero at infinity! That is, since $x > 0$ in the region of interest, then

$$\lim_{s \rightarrow \infty} \frac{1}{s^2 + 1} e^{sx} = \infty \neq 0.$$

It almost looks as if we could use the shifting property, but notice that the shift is in the wrong direction.

Of course, we need not restrict ourselves to first order equations, although the computations become more involved for higher order equations.

Example 6.5.3: Let us use Laplace for the following problem:

$$\begin{aligned} y_t &= y_{xx}, \quad 0 < x < \infty, \quad t > 0, \\ y_x(0, t) &= f(t), \\ y(x, 0) &= 0. \end{aligned}$$

Really we also impose other conditions on the solution so that for example the Laplace transform exists. For example, we might impose that y is bounded for each fixed time t .

Solution: Transform the equation in the t variable to find

$$sY(x) = Y''(x).$$

The general solution to this ODE is

$$Y(x) = Ae^{\sqrt{s}x} + Be^{-\sqrt{s}x}.$$

First $A = 0$, since otherwise Y does not decay to zero as $s \rightarrow \infty$.

Now consider the boundary condition. Transform $Y'(0) = F(s)$ and so $-\sqrt{s}B = F(s)$. In other words,

$$Y(x) = -F(s) \frac{1}{\sqrt{s}} e^{-\sqrt{s}x}.$$

If we look up the inverse transform in a table such as the one in [Appendix C](#) (or we spend the afternoon doing calculus), we find

$$\mathcal{L}^{-1} \left[e^{-\sqrt{s}x} \right] = \frac{x}{\sqrt{4\pi t^3}} e^{\frac{-x^2}{4t}},$$

or

$$\mathcal{L}^{-1} \left[\frac{1}{\sqrt{s}} e^{-\sqrt{s}x} \right] = \frac{1}{\sqrt{\pi t}} e^{\frac{-x^2}{4t}}.$$

So

$$y(x, t) = \mathcal{L}^{-1} \left[F(s) e^{-\sqrt{s}x} \right] = \int_0^t f(\tau) \frac{1}{\sqrt{\pi(t-\tau)}} e^{\frac{-x^2}{4(t-\tau)}} d\tau.$$

—

Laplace can solve problems where separation of variables fails. Laplace does not mind nonhomogeneity, but it is essentially only useful for constant coefficient equations.

6.5.1 Exercises

Exercise 6.5.1: Solve

$$\begin{aligned} y_t + y_x &= 1, \quad 0 < x < \infty, \quad t > 0, \\ y(0, t) &= 1, \quad y(x, 0) = 0. \end{aligned}$$

Exercise 6.5.2:* Solve

$$\begin{aligned} y_t + y_x &= 1, \quad 0 < x < \infty, \quad t > 0, \\ y(0, t) &= 0, \quad y(x, 0) = 0. \end{aligned}$$

Exercise 6.5.3: Solve

$$\begin{aligned} y_t + \alpha y_x &= 0, & 0 < x < \infty, t > 0, \\ y(0, t) &= t, & y(x, 0) = 0. \end{aligned}$$

Exercise 6.5.4: Solve

$$\begin{aligned} y_t + 2y_x &= x + t, & 0 < x < \infty, t > 0, \\ y(0, t) &= 0, & y(x, 0) = 0. \end{aligned}$$

Exercise 6.5.5: For an $\alpha > 0$, solve

$$\begin{aligned} y_t + \alpha y_x + y &= 0, & 0 < x < \infty, t > 0, \\ y(0, t) &= \sin(t), & y(x, 0) = 0. \end{aligned}$$

Exercise 6.5.6:* For a $c > 0$, solve

$$\begin{aligned} y_t + y_x + cy &= 0, & 0 < x < \infty, t > 0, \\ y(0, t) &= \sin(t), & y(x, 0) = 0. \end{aligned}$$

Exercise 6.5.7: Find the corresponding ODE problem for $Y(x)$, after transforming the t variable

$$\begin{aligned} y_{tt} + 3y_{xx} + y_{xt} + 3y_x + y &= \sin(x) + t, & 0 < x < 1, t > 0, \\ y(0, t) &= 1, & y(1, t) = t, & y(x, 0) = 1 - x, & y_t(x, 0) = 1. \end{aligned}$$

Do not solve the problem.

Exercise 6.5.8:* Find the corresponding ODE problem for $Y(x)$, after transforming the t variable

$$\begin{aligned} y_{tt} + 3y_{xx} + y &= x + t, & -1 < x < 1, t > 0, \\ y(-1, t) &= 0, & y(1, t) = 0, & y(x, 0) = (1 - x^2), & y_t(x, 0) = 0. \end{aligned}$$

Do not solve the problem.

Exercise 6.5.9: Write down a solution to

$$\begin{aligned} y_t &= y_{xx}, & 0 < x < \infty, t > 0, \\ y_x(0, t) &= e^{-t}, & y(x, 0) = 0, \end{aligned}$$

as an definite integral (convolution).

Exercise 6.5.10: Use the Laplace transform in t to solve

$$\begin{aligned} y_{tt} &= y_{xx}, & -\infty < x < \infty, t > 0, \\ y_t(x, 0) &= \sin(x), & y(x, 0) = 0. \end{aligned}$$

Hint: Note that e^{sx} does not go to zero as $s \rightarrow \infty$ for positive x , and e^{-sx} does not go to zero as $s \rightarrow \infty$ for negative x .

Exercise 6.5.11:* Use the Laplace transform in t to solve

$$\begin{aligned}y_{tt} &= y_{xx}, \quad -\infty < x < \infty, \quad t > 0, \\y_t(x, 0) &= x^2, \quad y(x, 0) = 0.\end{aligned}$$

Hint: Note that e^{sx} does not go to zero as $s \rightarrow \infty$ for positive x , and e^{-sx} does not go to zero as $s \rightarrow \infty$ for negative x .

Chapter 7

Power series methods

7.1 Power series

Attribution: [JL], §7.1.

Learning Objectives

After this section, you will be able to:

- Determine intervals of convergence for power series,
- Use differentiation and integration operations on power series, and
- Determine power series representations for rational functions.

Many functions can be written in terms of a power series

$$\sum_{k=0}^{\infty} a_k(x - x_0)^k.$$

If we assume that a solution of a differential equation is written as a power series, then perhaps we can use a method reminiscent of undetermined coefficients. That is, we will try to solve for the numbers a_k . Before we can carry out this process, let us review some results and concepts about power series.

7.1.1 Definition

Definition 7.1.1

A *power series* is an expression such as

$$\sum_{k=0}^{\infty} a_k(x - x_0)^k = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + a_3(x - x_0)^3 + \dots, \quad (7.1)$$

where $a_0, a_1, a_2, \dots, a_k, \dots$ and x_0 are constants.

Let

$$S_n(x) = \sum_{k=0}^n a_k(x - x_0)^k = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + a_3(x - x_0)^3 + \cdots + a_n(x - x_0)^n,$$

denote the so-called *partial sum*. If for some x , the limit

$$\lim_{n \rightarrow \infty} S_n(x) = \lim_{n \rightarrow \infty} \sum_{k=0}^n a_k(x - x_0)^k$$

exists, then we say that the series (7.1) *converges* at x . At $x = x_0$, the series always converges to a_0 . When (7.1) converges at any other point $x \neq x_0$, we say that (7.1) is a *convergent power series*, and we write

$$\sum_{k=0}^{\infty} a_k(x - x_0)^k = \lim_{n \rightarrow \infty} \sum_{k=0}^n a_k(x - x_0)^k.$$

If the series does not converge for any point $x \neq x_0$, we say that the series is *divergent*.

Example 7.1.1: The series

$$\sum_{k=0}^{\infty} \frac{1}{k!} x^k = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots$$

is convergent for any x . Recall that $k! = 1 \cdot 2 \cdot 3 \cdots k$ is the factorial. By convention we define $0! = 1$. You may recall that this series converges to e^x .

We say that (7.1) *converges absolutely* at x whenever the limit

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n |a_k| |x - x_0|^k$$

exists. That is, the series $\sum_{k=0}^{\infty} |a_k| |x - x_0|^k$ is convergent. If (7.1) converges absolutely at x , then it converges at x . However, the opposite implication is not true.

Example 7.1.2: The series

$$\sum_{k=1}^{\infty} \frac{1}{k} x^k$$

converges absolutely for all x in the interval $(-1, 1)$. It converges at $x = -1$, as $\sum_{k=1}^{\infty} \frac{(-1)^k}{k}$ converges (conditionally) by the alternating series test. The power series does not converge absolutely at $x = 1$, because $\sum_{k=1}^{\infty} \frac{1}{k}$ does not converge. The series diverges at $x = 1$.

7.1.2 Radius of convergence

If a power series converges absolutely at some x_1 , then for all x such that $|x - x_0| \leq |x_1 - x_0|$ (that is, x is closer than x_1 to x_0) we have $|a_k(x - x_0)^k| \leq |a_k(x_1 - x_0)^k|$ for all k . As

the numbers $|a_k(x_1 - x_0)^k|$ sum to some finite limit, summing smaller positive numbers $|a_k(x - x_0)^k|$ must also have a finite limit. Hence, the series must converge absolutely at x .

Theorem 7.1.1

For a power series (7.1), there exists a number ρ (we allow $\rho = \infty$) called the *radius of convergence* such that the series converges absolutely on the interval $(x_0 - \rho, x_0 + \rho)$ and diverges for $x < x_0 - \rho$ and $x > x_0 + \rho$. We write $\rho = \infty$ if the series converges for all x .

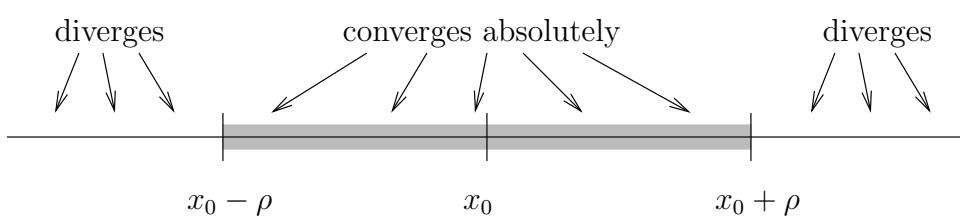


Figure 7.1: Convergence of a power series.

See Figure 7.1. In Example 7.1.1 the radius of convergence is $\rho = \infty$ as the series converges everywhere. In Example 7.1.2 the radius of convergence is $\rho = 1$. We note that $\rho = 0$ is another way of saying that the series is divergent.

A useful test for convergence of a series is the *ratio test*. Suppose that

$$\sum_{k=0}^{\infty} c_k$$

is a series and the limit

$$L = \lim_{n \rightarrow \infty} \left| \frac{c_{k+1}}{c_k} \right|$$

exists. Then the series converges absolutely if $L < 1$ and diverges if $L > 1$.

We apply this test to the series (7.1). Let $c_k = a_k(x - x_0)^k$ in the test. Compute

$$L = \lim_{n \rightarrow \infty} \left| \frac{c_{k+1}}{c_k} \right| = \lim_{n \rightarrow \infty} \left| \frac{a_{k+1}(x - x_0)^{k+1}}{a_k(x - x_0)^k} \right| = \lim_{n \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| |x - x_0|.$$

Define A by

$$A = \lim_{n \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right|.$$

Then if $1 > L = A|x - x_0|$ the series (7.1) converges absolutely. If $A = 0$, then the series always converges. If $A > 0$, then the series converges absolutely if $|x - x_0| < 1/A$, and diverges if $|x - x_0| > 1/A$. That is, the radius of convergence is $1/A$.

A similar test is the *root test*. Suppose

$$L = \lim_{k \rightarrow \infty} \sqrt[k]{|c_k|}$$

exists. Then $\sum_{k=0}^{\infty} c_k$ converges absolutely if $L < 1$ and diverges if $L > 1$. We can use the same calculation as above to find A . Let us summarize.

Theorem 7.1.2 (Ratio and root tests for power series)

Consider a power series

$$\sum_{k=0}^{\infty} a_k(x - x_0)^k$$

such that

$$A = \lim_{n \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| \quad \text{or} \quad A = \lim_{k \rightarrow \infty} \sqrt[k]{|a_k|}$$

exists. If $A = 0$, then the radius of convergence of the series is ∞ . Otherwise, the radius of convergence is $1/A$.

Example 7.1.3: Find the radius of convergence for the series

$$\sum_{k=0}^{\infty} 2^{-k}(x - 1)^k.$$

Solution: First we compute the limit in the ratio test,

$$A = \lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| = \lim_{k \rightarrow \infty} \left| \frac{2^{-k-1}}{2^{-k}} \right| = \lim_{k \rightarrow \infty} 2^{-1} = 1/2.$$

Therefore the radius of convergence is 2, and the series converges absolutely on the interval $(-1, 3)$. And we could just as well have used the root test:

$$A = \lim_{k \rightarrow \infty} \lim_{k \rightarrow \infty} \sqrt[k]{|a_k|} = \lim_{k \rightarrow \infty} \sqrt[k]{|2^{-k}|} = \lim_{k \rightarrow \infty} 2^{-1} = 1/2.$$

Example 7.1.4: Where does the series below converge?

$$\sum_{k=0}^{\infty} \frac{1}{k^k} x^k.$$

Solution: Compute the limit for the root test,

$$A = \lim_{k \rightarrow \infty} \sqrt[k]{|a_k|} = \lim_{k \rightarrow \infty} \sqrt[k]{\left| \frac{1}{k^k} \right|} = \lim_{k \rightarrow \infty} \sqrt[k]{\left| \frac{1}{k} \right|^k} = \lim_{k \rightarrow \infty} \frac{1}{k} = 0.$$

So the radius of convergence is ∞ : the series converges everywhere. The ratio test would also work here.

The root or the ratio test does not always apply. That is the limit of $\left| \frac{a_{k+1}}{a_k} \right|$ or $\sqrt[k]{|a_k|}$ might not exist. There exist more sophisticated ways of finding the radius of convergence, but those would be beyond the scope of this chapter. The two methods above cover many of the series that arise in practice. Often if the root test applies, so does the ratio test, and vice versa, though the limit might be easier to compute in one way than the other.

7.1.3 Analytic functions

Functions represented by power series are called *analytic functions*. Not every function is analytic, although the majority of the functions you have seen in calculus are.

An analytic function $f(x)$ is equal to its *Taylor series** near a point x_0 . That is, for x near x_0 we have

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k, \quad (7.2)$$

where $f^{(k)}(x_0)$ denotes the k^{th} derivative of $f(x)$ at the point x_0 .

For example, sine is an analytic function and its Taylor series around $x_0 = 0$ is given by

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}.$$

In Figure 7.2 we plot $\sin(x)$ and the truncations of the series up to degree 5 and 9. You can see that the approximation is very good for x near 0, but gets worse for x further away from 0. This is what happens in general. To get a good approximation far away from x_0 you need to take more and more terms of the Taylor series.

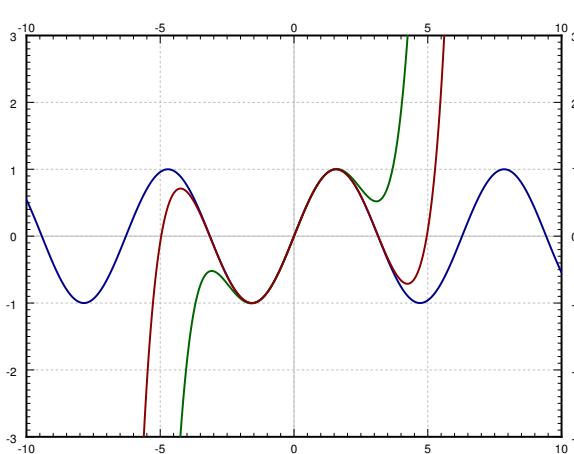


Figure 7.2: The sine function and its Taylor approximations around $x_0 = 0$ of 5th and 9th degree.

7.1.4 Manipulating power series

One of the main properties of power series that we will use is that we can differentiate them term by term. That is, suppose that $\sum a_k(x - x_0)^k$ is a convergent power series. Then for x in the radius of convergence we have

$$\frac{d}{dx} \left[\sum_{k=0}^{\infty} a_k (x - x_0)^k \right] = \sum_{k=1}^{\infty} k a_k (x - x_0)^{k-1}.$$

*Named after the English mathematician Sir Brook Taylor (1685–1731).

Notice that the term corresponding to $k = 0$ disappeared as it was constant. The radius of convergence of the differentiated series is the same as that of the original.

Example 7.1.5: Show that the exponential $y = e^x$ solves $y' = y$ using power series.

Solution: First write

$$y = e^x = \sum_{k=0}^{\infty} \frac{1}{k!} x^k.$$

Now differentiate

$$y' = \sum_{k=1}^{\infty} k \frac{1}{k!} x^{k-1} = \sum_{k=1}^{\infty} \frac{1}{(k-1)!} x^{k-1}.$$

We *reindex* the series by simply replacing k with $k + 1$. The series does not change, what changes is simply how we write it. After reindexing the series starts at $k = 0$ again.

$$\sum_{k=1}^{\infty} \frac{1}{(k-1)!} x^{k-1} = \sum_{k+1=1}^{\infty} \frac{1}{((k+1)-1)!} x^{(k+1)-1} = \sum_{k=0}^{\infty} \frac{1}{k!} x^k.$$

That was precisely the power series for e^x that we started with, so we showed that $\frac{d}{dx}[e^x] = e^x$.

Convergent power series can be added and multiplied together, and multiplied by constants using the following rules. First, we can add series by adding term by term,

$$\left(\sum_{k=0}^{\infty} a_k (x - x_0)^k \right) + \left(\sum_{k=0}^{\infty} b_k (x - x_0)^k \right) = \sum_{k=0}^{\infty} (a_k + b_k) (x - x_0)^k.$$

We can multiply by constants,

$$\alpha \left(\sum_{k=0}^{\infty} a_k (x - x_0)^k \right) = \sum_{k=0}^{\infty} \alpha a_k (x - x_0)^k.$$

We can also multiply series together,

$$\left(\sum_{k=0}^{\infty} a_k (x - x_0)^k \right) \left(\sum_{k=0}^{\infty} b_k (x - x_0)^k \right) = \sum_{k=0}^{\infty} c_k (x - x_0)^k,$$

where $c_k = a_0 b_k + a_1 b_{k-1} + \cdots + a_k b_0$. The radius of convergence of the sum or the product is at least the minimum of the radii of convergence of the two series involved.

7.1.5 Power series for rational functions

Polynomials are simply finite power series. That is, a polynomial is a power series where the a_k are zero for all k large enough. We can always expand a polynomial as a power series about any point x_0 by writing the polynomial as a polynomial in $(x - x_0)$. For example, let us write $2x^2 - 3x + 4$ as a power series around $x_0 = 1$:

$$2x^2 - 3x + 4 = 3 + (x - 1) + 2(x - 1)^2.$$

In other words $a_0 = 3$, $a_1 = 1$, $a_2 = 2$, and all other $a_k = 0$. To do this, we know that $a_k = 0$ for all $k \geq 3$ as the polynomial is of degree 2. We write $a_0 + a_1(x - 1) + a_2(x - 1)^2$, we expand, and we solve for a_0 , a_1 , and a_2 . We could have also differentiated at $x = 1$ and used the Taylor series formula (7.2).

Let us look at rational functions, that is, ratios of polynomials. An important fact is that a series for a function only defines the function on an interval even if the function is defined elsewhere. For example, for $-1 < x < 1$ we have

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + \dots$$

This series is called the *geometric series*. The ratio test tells us that the radius of convergence is 1. The series diverges for $x \leq -1$ and $x \geq 1$, even though $\frac{1}{1-x}$ is defined for all $x \neq 1$.

We can use the geometric series together with rules for addition and multiplication of power series to expand rational functions around a point, as long as the denominator is not zero at x_0 . Note that as for polynomials, we could equivalently use the Taylor series expansion (7.2).

Example 7.1.6: Expand $\frac{x}{1+2x+x^2}$ as a power series around the origin ($x_0 = 0$) and find the radius of convergence.

Solution: First, write $1 + 2x + x^2 = (1 + x)^2 = (1 - (-x))^2$. Compute

$$\begin{aligned} \frac{x}{1+2x+x^2} &= x \left(\frac{1}{1-(-x)} \right)^2 \\ &= x \left(\sum_{k=0}^{\infty} (-1)^k x^k \right)^2 \\ &= x \left(\sum_{k=0}^{\infty} c_k x^k \right) \\ &= \sum_{k=0}^{\infty} c_k x^{k+1}, \end{aligned}$$

where to get c_k , we use the formula for the product of series. We obtain, $c_0 = 1$, $c_1 = -1 - 1 = -2$, $c_2 = 1 + 1 + 1 = 3$, etc. Therefore

$$\frac{x}{1+2x+x^2} = \sum_{k=1}^{\infty} (-1)^{k+1} k x^k = x - 2x^2 + 3x^3 - 4x^4 + \dots$$

The radius of convergence is at least 1. We use the ratio test

$$\lim_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| = \lim_{k \rightarrow \infty} \left| \frac{(-1)^{k+2}(k+1)}{(-1)^{k+1}k} \right| = \lim_{k \rightarrow \infty} \frac{k+1}{k} = 1.$$

So the radius of convergence is actually equal to 1. □

When the rational function is more complicated, it is also possible to use method of partial fractions. For example, to find the Taylor series for $\frac{x^3+x}{x^2-1}$, we write

$$\frac{x^3+x}{x^2-1} = x + \frac{1}{1+x} - \frac{1}{1-x} = x + \sum_{k=0}^{\infty} (-1)^k x^k - \sum_{k=0}^{\infty} x^k = -x + \sum_{\substack{k=3 \\ k \text{ odd}}}^{\infty} (-2)x^k.$$

7.1.6 Exercises

Exercise 7.1.1: Is the power series $\sum_{k=0}^{\infty} e^k x^k$ convergent? If so, what is the radius of convergence?

Exercise 7.1.2: Is the power series $\sum_{k=0}^{\infty} kx^k$ convergent? If so, what is the radius of convergence?

Exercise 7.1.3:* Is the power series $\sum_{n=1}^{\infty} (0.1)^n x^n$ convergent? If so, what is the radius of convergence?

Exercise 7.1.4: Is the power series $\sum_{k=0}^{\infty} k! x^k$ convergent? If so, what is the radius of convergence?

Exercise 7.1.5: Is the power series $\sum_{k=0}^{\infty} \frac{1}{(2k)!} (x-10)^k$ convergent? If so, what is the radius of convergence?

Exercise 7.1.6 (challenging):* Is the power series $\sum_{n=1}^{\infty} \frac{n!}{n^n} x^n$ convergent? If so, what is the radius of convergence?

Exercise 7.1.7: Determine the Taylor series for $\sin x$ around the point $x_0 = \pi$.

Exercise 7.1.8: Determine the Taylor series for $\ln x$ around the point $x_0 = 1$, and find the radius of convergence.

Exercise 7.1.9 (challenging):* Find the Taylor series for $x^7 e^x$ around $x_0 = 0$.

Exercise 7.1.10: Determine the Taylor series and its radius of convergence of $\frac{1}{1+x}$ around $x_0 = 0$.

Exercise 7.1.11: Determine the Taylor series and its radius of convergence of $\frac{x}{4-x^2}$ around $x_0 = 0$. Hint: You will not be able to use the ratio test.

Exercise 7.1.12: Expand $x^5 + 5x + 1$ as a power series around $x_0 = 5$.

Exercise 7.1.13:* Using the geometric series, expand $\frac{1}{1-x}$ around $x_0 = 2$. For what x does the series converge?

Exercise 7.1.14: Suppose that the ratio test applies to a series $\sum_{k=0}^{\infty} a_k x^k$. Show, using the ratio test, that the radius of convergence of the differentiated series is the same as that of the original series.

Exercise 7.1.15: Suppose that f is an analytic function such that $f^{(n)}(0) = n$. Find $f(1)$.

Exercise 7.1.16 (challenging);* Imagine f and g are analytic functions such that $f^{(k)}(0) = g^{(k)}(0)$ for all large enough k . What can you say about $f(x) - g(x)$?

7.2 Series solutions of linear second order ODEs

Attribution: [JL], §7.2.

Learning Objectives

After this section, you will be able to:

- Use power series methods to solve second order linear ODEs near ordinary points and
- Write a recurrence relation for the coefficients in a power series solution to an ODE.

Suppose we have a linear second order homogeneous ODE of the form

$$p(x)y'' + q(x)y' + r(x)y = 0.$$

Suppose that $p(x)$, $q(x)$, and $r(x)$ are polynomials. We will try a solution of the form

$$y = \sum_{k=0}^{\infty} a_k(x - x_0)^k$$

and solve for the a_k to try to obtain a solution defined in some interval around x_0 .

The point x_0 is called an *ordinary point* if $p(x_0) \neq 0$. That is, the functions

$$\frac{q(x)}{p(x)} \quad \text{and} \quad \frac{r(x)}{p(x)}$$

are defined for x near x_0 . If $p(x_0) = 0$, then we say x_0 is a *singular point*. Handling singular points is harder than ordinary points and so we now focus only on ordinary points.

Example 7.2.1: Let us start with a very simple example

$$y'' - y = 0.$$

Solution: Let us try a power series solution near $x_0 = 0$, which is an ordinary point. Every point is an ordinary point in fact, as the equation is constant coefficient. We already know we should obtain exponentials or the hyperbolic sine and cosine, but let us pretend we do not know this.

We try

$$y = \sum_{k=0}^{\infty} a_k x^k.$$

If we differentiate, the $k = 0$ term is a constant and hence disappears. We therefore get

$$y' = \sum_{k=1}^{\infty} k a_k x^{k-1}.$$

We differentiate yet again to obtain (now the $k = 1$ term disappears)

$$y'' = \sum_{k=2}^{\infty} k(k-1)a_k x^{k-2}.$$

We reindex the series (replace k with $k+2$) to obtain

$$y'' = \sum_{k=0}^{\infty} (k+2)(k+1) a_{k+2} x^k.$$

Now we plug y and y'' into the differential equation

$$\begin{aligned} 0 = y'' - y &= \left(\sum_{k=0}^{\infty} (k+2)(k+1) a_{k+2} x^k \right) - \left(\sum_{k=0}^{\infty} a_k x^k \right) \\ &= \sum_{k=0}^{\infty} ((k+2)(k+1) a_{k+2} x^k - a_k x^k) \\ &= \sum_{k=0}^{\infty} ((k+2)(k+1) a_{k+2} - a_k) x^k. \end{aligned}$$

As $y'' - y$ is supposed to be equal to 0, we know that the coefficients of the resulting series must be equal to 0. Therefore,

$$(k+2)(k+1) a_{k+2} - a_k = 0, \quad \text{or} \quad a_{k+2} = \frac{a_k}{(k+2)(k+1)}.$$

The equation above is called a *recurrence relation* for the coefficients of the power series. It did not matter what a_0 or a_1 was. They can be arbitrary. But once we pick a_0 and a_1 , then all other coefficients are determined by the recurrence relation.

Let us see what the coefficients must be. First, a_0 and a_1 are arbitrary. Then,

$$a_2 = \frac{a_0}{2}, \quad a_3 = \frac{a_1}{(3)(2)}, \quad a_4 = \frac{a_2}{(4)(3)} = \frac{a_0}{(4)(3)(2)}, \quad a_5 = \frac{a_3}{(5)(4)} = \frac{a_1}{(5)(4)(3)(2)}, \quad \dots$$

So for even k , that is $k = 2n$, we have

$$a_k = a_{2n} = \frac{a_0}{(2n)!},$$

and for odd k , that is $k = 2n+1$, we have

$$a_k = a_{2n+1} = \frac{a_1}{(2n+1)!}.$$

Let us write down the series

$$y = \sum_{k=0}^{\infty} a_k x^k = \sum_{n=0}^{\infty} \left(\frac{a_0}{(2n)!} x^{2n} + \frac{a_1}{(2n+1)!} x^{2n+1} \right) = a_0 \sum_{n=0}^{\infty} \frac{1}{(2n)!} x^{2n} + a_1 \sum_{n=0}^{\infty} \frac{1}{(2n+1)!} x^{2n+1}.$$

We recognize the two series as the hyperbolic sine and cosine. Therefore,

$$y = a_0 \cosh x + a_1 \sinh x.$$

□

Of course, in general we will not be able to recognize the series that appears, since usually there will not be any elementary function that matches it. In that case we will be content with the series.

Example 7.2.2: Let us do a more complex example. Consider *Airy's equation*^{*}:

$$y'' - xy = 0,$$

near the point $x_0 = 0$. Note that $x_0 = 0$ is an ordinary point.

Solution: We try

$$y = \sum_{k=0}^{\infty} a_k x^k.$$

We differentiate twice (as above) to obtain

$$y'' = \sum_{k=2}^{\infty} k(k-1) a_k x^{k-2}.$$

We plug y into the equation

$$\begin{aligned} 0 = y'' - xy &= \left(\sum_{k=2}^{\infty} k(k-1) a_k x^{k-2} \right) - x \left(\sum_{k=0}^{\infty} a_k x^k \right) \\ &= \left(\sum_{k=2}^{\infty} k(k-1) a_k x^{k-2} \right) - \left(\sum_{k=0}^{\infty} a_k x^{k+1} \right). \end{aligned}$$

We reindex to make things easier to sum

$$\begin{aligned} 0 = y'' - xy &= \left(2a_2 + \sum_{k=1}^{\infty} (k+2)(k+1) a_{k+2} x^k \right) - \left(\sum_{k=1}^{\infty} a_{k-1} x^k \right) \\ &= 2a_2 + \sum_{k=1}^{\infty} ((k+2)(k+1) a_{k+2} - a_{k-1}) x^k. \end{aligned}$$

Again $y'' - xy$ is supposed to be 0, so $a_2 = 0$, and

$$(k+2)(k+1) a_{k+2} - a_{k-1} = 0, \quad \text{or} \quad a_{k+2} = \frac{a_{k-1}}{(k+2)(k+1)}.$$

We jump in steps of three. First, since $a_2 = 0$ we must have $a_5 = 0, a_8 = 0, a_{11} = 0$, etc. In general, $a_{3n+2} = 0$.

*Named after the English mathematician Sir George Biddell Airy (1801–1892).

The constants a_0 and a_1 are arbitrary and we obtain

$$a_3 = \frac{a_0}{(3)(2)}, \quad a_4 = \frac{a_1}{(4)(3)}, \quad a_6 = \frac{a_3}{(6)(5)} = \frac{a_0}{(6)(5)(3)(2)}, \quad a_7 = \frac{a_4}{(7)(6)} = \frac{a_1}{(7)(6)(4)(3)}, \quad \dots$$

For a_k where k is a multiple of 3, that is $k = 3n$ we notice that

$$a_{3n} = \frac{a_0}{(2)(3)(5)(6) \cdots (3n-1)(3n)}.$$

For a_k where $k = 3n + 1$, we notice

$$a_{3n+1} = \frac{a_1}{(3)(4)(6)(7) \cdots (3n)(3n+1)}.$$

In other words, if we write down the series for y , it has two parts

$$\begin{aligned} y &= \left(a_0 + \frac{a_0}{6}x^3 + \frac{a_0}{180}x^6 + \cdots + \frac{a_0}{(2)(3)(5)(6) \cdots (3n-1)(3n)}x^{3n} + \cdots \right) \\ &\quad + \left(a_1x + \frac{a_1}{12}x^4 + \frac{a_1}{504}x^7 + \cdots + \frac{a_1}{(3)(4)(6)(7) \cdots (3n)(3n+1)}x^{3n+1} + \cdots \right) \\ &= a_0 \left(1 + \frac{1}{6}x^3 + \frac{1}{180}x^6 + \cdots + \frac{1}{(2)(3)(5)(6) \cdots (3n-1)(3n)}x^{3n} + \cdots \right) \\ &\quad + a_1 \left(x + \frac{1}{12}x^4 + \frac{1}{504}x^7 + \cdots + \frac{1}{(3)(4)(6)(7) \cdots (3n)(3n+1)}x^{3n+1} + \cdots \right). \end{aligned}$$

We define

$$\begin{aligned} y_1(x) &= 1 + \frac{1}{6}x^3 + \frac{1}{180}x^6 + \cdots + \frac{1}{(2)(3)(5)(6) \cdots (3n-1)(3n)}x^{3n} + \cdots, \\ y_2(x) &= x + \frac{1}{12}x^4 + \frac{1}{504}x^7 + \cdots + \frac{1}{(3)(4)(6)(7) \cdots (3n)(3n+1)}x^{3n+1} + \cdots, \end{aligned}$$

and write the general solution to the equation as $y(x) = a_0y_1(x) + a_1y_2(x)$. If we plug in $x = 0$ into the power series for y_1 and y_2 , we find $y_1(0) = 1$ and $y_2(0) = 0$. Similarly, $y'_1(0) = 0$ and $y'_2(0) = 1$. Therefore $y = a_0y_1 + a_1y_2$ is a solution that satisfies the initial conditions $y(0) = a_0$ and $y'(0) = a_1$. \square

The functions y_1 and y_2 cannot be written in terms of the elementary functions that you know. See [Figure 7.3](#) on the next page for the plot of the solutions y_1 and y_2 . These functions have many interesting properties. For example, they are oscillatory for negative x (like solutions to $y'' + y = 0$) and for positive x they grow without bound (like solutions to $y'' - y = 0$).

Sometimes a solution may turn out to be a polynomial.

Example 7.2.3: Find a solution to the so-called *Hermite's equation of order n* ^{*}:

$$y'' - 2xy' + 2ny = 0.$$

*Named after the French mathematician [Charles Hermite](#) (1822–1901).

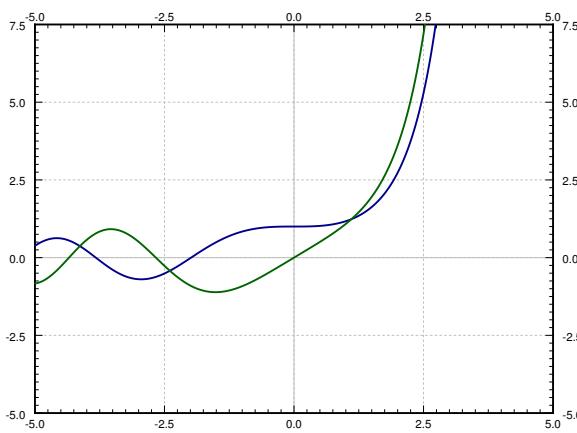


Figure 7.3: The two solutions y_1 and y_2 to Airy's equation.

Solution: Let us find a solution around the point $x_0 = 0$. We try

$$y = \sum_{k=0}^{\infty} a_k x^k.$$

We differentiate (as above) to obtain

$$\begin{aligned} y' &= \sum_{k=1}^{\infty} k a_k x^{k-1}, \\ y'' &= \sum_{k=2}^{\infty} k(k-1) a_k x^{k-2}. \end{aligned}$$

Now we plug into the equation

$$\begin{aligned} 0 &= y'' - 2xy' + 2ny \\ &= \left(\sum_{k=2}^{\infty} k(k-1) a_k x^{k-2} \right) - 2x \left(\sum_{k=1}^{\infty} k a_k x^{k-1} \right) + 2n \left(\sum_{k=0}^{\infty} a_k x^k \right) \\ &= \left(\sum_{k=2}^{\infty} k(k-1) a_k x^{k-2} \right) - \left(\sum_{k=1}^{\infty} 2ka_k x^k \right) + \left(\sum_{k=0}^{\infty} 2na_k x^k \right) \\ &= \left(2a_2 + \sum_{k=1}^{\infty} (k+2)(k+1)a_{k+2} x^k \right) - \left(\sum_{k=1}^{\infty} 2ka_k x^k \right) + \left(2na_0 + \sum_{k=1}^{\infty} 2na_k x^k \right) \\ &= 2a_2 + 2na_0 + \sum_{k=1}^{\infty} ((k+2)(k+1)a_{k+2} - 2ka_k + 2na_k) x^k. \end{aligned}$$

As $y'' - 2xy' + 2ny = 0$ we have

$$(k+2)(k+1)a_{k+2} + (-2k+2n)a_k = 0, \quad \text{or} \quad a_{k+2} = \frac{(2k-2n)}{(k+2)(k+1)} a_k.$$

This recurrence relation actually includes $a_2 = -na_0$ (which comes about from $2a_2 + 2na_0 = 0$). Again a_0 and a_1 are arbitrary.

$$\begin{aligned} a_2 &= \frac{-2n}{(2)(1)}a_0, & a_3 &= \frac{2(1-n)}{(3)(2)}a_1, \\ a_4 &= \frac{2(2-n)}{(4)(3)}a_2 = \frac{2^2(2-n)(-n)}{(4)(3)(2)(1)}a_0, \\ a_5 &= \frac{2(3-n)}{(5)(4)}a_3 = \frac{2^2(3-n)(1-n)}{(5)(4)(3)(2)}a_1, \quad \dots \end{aligned}$$

Let us separate the even and odd coefficients. We find that

$$\begin{aligned} a_{2m} &= \frac{2^m(-n)(2-n)\cdots(2m-2-n)}{(2m)!}, \\ a_{2m+1} &= \frac{2^m(1-n)(3-n)\cdots(2m-1-n)}{(2m+1)!}. \end{aligned}$$

Let us write down the two series, one with the even powers and one with the odd.

$$\begin{aligned} y_1(x) &= 1 + \frac{2(-n)}{2!}x^2 + \frac{2^2(-n)(2-n)}{4!}x^4 + \frac{2^3(-n)(2-n)(4-n)}{6!}x^6 + \dots, \\ y_2(x) &= x + \frac{2(1-n)}{3!}x^3 + \frac{2^2(1-n)(3-n)}{5!}x^5 + \frac{2^3(1-n)(3-n)(5-n)}{7!}x^7 + \dots. \end{aligned}$$

We then write

$$y(x) = a_0y_1(x) + a_1y_2(x).$$

We remark that if n is a positive even integer, then $y_1(x)$ is a polynomial as all the coefficients in the series beyond a certain degree are zero. If n is a positive odd integer, then $y_2(x)$ is a polynomial. For example, if $n = 4$, then

$$y_1(x) = 1 + \frac{2(-4)}{2!}x^2 + \frac{2^2(-4)(2-4)}{4!}x^4 = 1 - 4x^2 + \frac{4}{3}x^4.$$

\(\square\)

7.2.1 Exercises

In the following exercises, when asked to solve an equation using power series methods, you should find the first few terms of the series, and if possible find a general formula for the k^{th} coefficient.

Exercise 7.2.1: Use power series methods to solve $y'' + y = 0$ at the point $x_0 = 1$.

Exercise 7.2.2: Use power series methods to solve $y'' + 4xy = 0$ at the point $x_0 = 0$.

Exercise 7.2.3:* Use power series methods to solve $y'' + 2x^3y = 0$ at the point $x_0 = 0$.

Exercise 7.2.4: Use power series methods to solve $y'' - xy = 0$ at the point $x_0 = 1$.

Exercise 7.2.5: Use power series methods to solve $y'' + x^2y = 0$ at the point $x_0 = 0$.

Exercise 7.2.6: The methods work for other orders than second order. Try the methods of this section to solve the first order system $y' - xy = 0$ at the point $x_0 = 0$.

Exercise 7.2.7:* Attempt to solve $x^2y'' - y = 0$ at $x_0 = 0$ using the power series method of this section (x_0 is a singular point). Can you find at least one solution? Can you find more than one solution?

Exercise 7.2.8 (Chebyshev's equation of order p):

- a) Solve $(1 - x^2)y'' - xy' + p^2y = 0$ using power series methods at $x_0 = 0$.
- b) For what p is there a polynomial solution?

Exercise 7.2.9: Find a polynomial solution to $(x^2 + 1)y'' - 2xy' + 2y = 0$ using power series methods.

Exercise 7.2.10:

- a) Use power series methods to solve $(1 - x)y'' + y = 0$ at the point $x_0 = 0$.
- b) Use the solution to part a) to find a solution for $xy'' + y = 0$ around the point $x_0 = 1$.

Exercise 7.2.11 (challenging):* Power series methods also work for nonhomogeneous equations.

- a) Use power series methods to solve $y'' - xy = \frac{1}{1-x}$ at the point $x_0 = 0$. Hint: Recall the geometric series.
- b) Now solve for the initial condition $y(0) = 0$, $y'(0) = 0$.

7.3 Singular points and the method of Frobenius

Attribution: [JL], §7.3.

Learning Objectives

After this section, you will be able to:

- Find power series solutions to an ODE at a singular point and
- Use the method of Frobenius to find solutions to ODE, including the Bessel equation.

While behavior of ODEs at singular points is more complicated, certain singular points are not especially difficult to solve. Let us look at some examples before giving a general method. We may be lucky and obtain a power series solution using the method of the previous section, but in general we may have to try other things.

7.3.1 Examples

Example 7.3.1: Let us first look at a simple first order equation

$$2xy' - y = 0.$$

Solution: Note that $x = 0$ is a singular point. If we try to plug in

$$y = \sum_{k=0}^{\infty} a_k x^k,$$

we obtain

$$\begin{aligned} 0 &= 2xy' - y = 2x \left(\sum_{k=1}^{\infty} ka_k x^{k-1} \right) - \left(\sum_{k=0}^{\infty} a_k x^k \right) \\ &= a_0 + \sum_{k=1}^{\infty} (2ka_k - a_k) x^k. \end{aligned}$$

First, $a_0 = 0$. Next, the only way to solve $0 = 2ka_k - a_k = (2k - 1)a_k$ for $k = 1, 2, 3, \dots$ is for $a_k = 0$ for all k . Therefore we only get the trivial solution $y = 0$. We need a nonzero solution to get the general solution.

Let us try $y = x^r$ for some real number r . Consequently our solution—if we can find one—may only make sense for positive x . Then $y' = rx^{r-1}$. So

$$0 = 2xy' - y = 2xrx^{r-1} - x^r = (2r - 1)x^r.$$

Therefore $r = 1/2$, or in other words $y = x^{1/2}$. Multiplying by a constant, the general solution for positive x is

$$y = Cx^{1/2}.$$

If $C \neq 0$, then the derivative of the solution “blows up” at $x = 0$ (the singular point). There is only one solution that is differentiable at $x = 0$ and that’s the trivial solution $y = 0$. \square

Not every problem with a singular point has a solution of the form $y = x^r$, of course. But perhaps we can combine the methods. What we will do is to try a solution of the form

$$y = x^r f(x)$$

where $f(x)$ is an analytic function.

Example 7.3.2: Consider the equation

$$4x^2y'' - 4x^2y' + (1 - 2x)y = 0,$$

and again note that $x = 0$ is a singular point. Attempt to solve this equation using the method above.

Solution: Let us try

$$y = x^r \sum_{k=0}^{\infty} a_k x^k = \sum_{k=0}^{\infty} a_k x^{k+r},$$

where r is a real number, not necessarily an integer. Again if such a solution exists, it may only exist for positive x . First let us find the derivatives

$$\begin{aligned} y' &= \sum_{k=0}^{\infty} (k+r) a_k x^{k+r-1}, \\ y'' &= \sum_{k=0}^{\infty} (k+r)(k+r-1) a_k x^{k+r-2}. \end{aligned}$$

Plugging into our equation we obtain

$$\begin{aligned}
0 &= 4x^2y'' - 4x^2y' + (1 - 2x)y \\
&= 4x^2 \left(\sum_{k=0}^{\infty} (k+r)(k+r-1) a_k x^{k+r-2} \right) - 4x^2 \left(\sum_{k=0}^{\infty} (k+r) a_k x^{k+r-1} \right) + (1 - 2x) \left(\sum_{k=0}^{\infty} a_k x^{k+r} \right) \\
&= \left(\sum_{k=0}^{\infty} 4(k+r)(k+r-1) a_k x^{k+r} \right) \\
&\quad - \left(\sum_{k=0}^{\infty} 4(k+r) a_k x^{k+r+1} \right) + \left(\sum_{k=0}^{\infty} a_k x^{k+r} \right) - \left(\sum_{k=0}^{\infty} 2a_k x^{k+r+1} \right) \\
&= \left(\sum_{k=0}^{\infty} 4(k+r)(k+r-1) a_k x^{k+r} \right) \\
&\quad - \left(\sum_{k=1}^{\infty} 4(k+r-1) a_{k-1} x^{k+r} \right) + \left(\sum_{k=0}^{\infty} a_k x^{k+r} \right) - \left(\sum_{k=1}^{\infty} 2a_{k-1} x^{k+r} \right) \\
&= 4r(r-1) a_0 x^r + a_0 x^r + \sum_{k=1}^{\infty} \left(4(k+r)(k+r-1) a_k - 4(k+r-1) a_{k-1} + a_k - 2a_{k-1} \right) x^{k+r} \\
&= (4r(r-1) + 1) a_0 x^r + \sum_{k=1}^{\infty} \left((4(k+r)(k+r-1) + 1) a_k - (4(k+r-1) + 2) a_{k-1} \right) x^{k+r}.
\end{aligned}$$

To have a solution we must first have $(4r(r-1) + 1) a_0 = 0$. Supposing that $a_0 \neq 0$ we obtain

$$4r(r-1) + 1 = 0.$$

This equation is called the *indicial equation*. This particular indicial equation has a double root at $r = 1/2$.

OK, so we know what r has to be. That knowledge we obtained simply by looking at the coefficient of x^r . All other coefficients of x^{k+r} also have to be zero so

$$(4(k+r)(k+r-1) + 1) a_k - (4(k+r-1) + 2) a_{k-1} = 0.$$

If we plug in $r = 1/2$ and solve for a_k , we get

$$a_k = \frac{4(k+1/2-1)+2}{4(k+1/2)(k+1/2-1)+1} a_{k-1} = \frac{1}{k} a_{k-1}.$$

Let us set $a_0 = 1$. Then

$$a_1 = \frac{1}{1} a_0 = 1, \quad a_2 = \frac{1}{2} a_1 = \frac{1}{2}, \quad a_3 = \frac{1}{3} a_2 = \frac{1}{3 \cdot 2}, \quad a_4 = \frac{1}{4} a_3 = \frac{1}{4 \cdot 3 \cdot 2}, \quad \dots$$

Extrapolating, we notice that

$$a_k = \frac{1}{k(k-1)(k-2)\cdots 3 \cdot 2} = \frac{1}{k!}.$$

In other words,

$$y = \sum_{k=0}^{\infty} a_k x^{k+r} = \sum_{k=0}^{\infty} \frac{1}{k!} x^{k+1/2} = x^{1/2} \sum_{k=0}^{\infty} \frac{1}{k!} x^k = x^{1/2} e^x.$$

That was lucky! In general, we will not be able to write the series in terms of elementary functions.

We have one solution, let us call it $y_1 = x^{1/2} e^x$. But what about a second solution? If we want a general solution, we need two linearly independent solutions. Picking a_0 to be a different constant only gets us a constant multiple of y_1 , and we do not have any other r to try; we only have one solution to the indicial equation. Well, there are powers of x floating around and we are taking derivatives, perhaps the logarithm (the antiderivative of x^{-1}) is around as well. It turns out we want to try for another solution of the form

$$y_2 = \sum_{k=0}^{\infty} b_k x^{k+r} + (\ln x) y_1,$$

which in our case is

$$y_2 = \sum_{k=0}^{\infty} b_k x^{k+1/2} + (\ln x) x^{1/2} e^x.$$

We now differentiate this equation, substitute into the differential equation and solve for b_k . A long computation ensues and we obtain some recursion relation for b_k . The reader can (and should) try this to obtain for example the first three terms

$$b_1 = b_0 - 1, \quad b_2 = \frac{2b_1 - 1}{4}, \quad b_3 = \frac{6b_2 - 1}{18}, \quad \dots$$

We then fix b_0 and obtain a solution y_2 . Then we write the general solution as $y = Ay_1 + B\underline{y_2}$.

7.3.2 The method of Frobenius

Before giving the general method, let us clarify when the method applies. Let

$$p(x)y'' + q(x)y' + r(x)y = 0$$

be an ODE. As before, if $p(x_0) = 0$, then x_0 is a singular point. If, furthermore, the limits

$$\lim_{x \rightarrow x_0} (x - x_0) \frac{q(x)}{p(x)} \quad \text{and} \quad \lim_{x \rightarrow x_0} (x - x_0)^2 \frac{r(x)}{p(x)}$$

both exist and are finite, then we say that x_0 is a *regular singular point*.

Example 7.3.3: Often, and for the rest of this section, $x_0 = 0$. Consider

$$x^2 y'' + x(1+x)y' + (\pi + x^2)y = 0.$$

Write

$$\lim_{x \rightarrow 0} x \frac{q(x)}{p(x)} = \lim_{x \rightarrow 0} x \frac{x(1+x)}{x^2} = \lim_{x \rightarrow 0} (1+x) = 1,$$

$$\lim_{x \rightarrow 0} x^2 \frac{r(x)}{p(x)} = \lim_{x \rightarrow 0} x^2 \frac{(\pi + x^2)}{x^2} = \lim_{x \rightarrow 0} (\pi + x^2) = \pi.$$

So $x = 0$ is a regular singular point.

On the other hand if we make the slight change

$$x^2 y'' + (1+x)y' + (\pi + x^2)y = 0,$$

then

$$\lim_{x \rightarrow 0} x \frac{q(x)}{p(x)} = \lim_{x \rightarrow 0} x \frac{(1+x)}{x^2} = \lim_{x \rightarrow 0} \frac{1+x}{x} = \text{DNE}.$$

Here DNE stands for *does not exist*. The point 0 is a singular point, but not a regular singular point.

Let us now discuss the general *Method of Frobenius*^{*}. We only consider the method at the point $x = 0$ for simplicity. The main idea is the following theorem.

Theorem 7.3.1 (Method of Frobenius)

Suppose that

$$p(x)y'' + q(x)y' + r(x)y = 0 \quad (7.3)$$

has a regular singular point at $x = 0$, then there exists at least one solution of the form

$$y = x^r \sum_{k=0}^{\infty} a_k x^k.$$

A solution of this form is called a *Frobenius-type solution*.

The method usually breaks down like this:

- (i) We seek a Frobenius-type solution of the form

$$y = \sum_{k=0}^{\infty} a_k x^{k+r}.$$

We plug this y into equation (7.3). We collect terms and write everything as a single series.

- (ii) The obtained series must be zero. Setting the first coefficient (usually the coefficient of x^r) in the series to zero we obtain the *indicial equation*, which is a quadratic polynomial in r .

*Named after the German mathematician Ferdinand Georg Frobenius (1849–1917).

- (iii) If the indicial equation has two real roots r_1 and r_2 such that $r_1 - r_2$ is not an integer, then we have two linearly independent Frobenius-type solutions. Using the first root, we plug in

$$y_1 = x^{r_1} \sum_{k=0}^{\infty} a_k x^k,$$

and we solve for all a_k to obtain the first solution. Then using the second root, we plug in

$$y_2 = x^{r_2} \sum_{k=0}^{\infty} b_k x^k,$$

and solve for all b_k to obtain the second solution.

- (iv) If the indicial equation has a doubled root r , then there we find one solution

$$y_1 = x^r \sum_{k=0}^{\infty} a_k x^k,$$

and then we obtain a new solution by plugging

$$y_2 = x^r \sum_{k=0}^{\infty} b_k x^k + (\ln x)y_1,$$

into equation (7.3) and solving for the constants b_k .

- (v) If the indicial equation has two real roots such that $r_1 - r_2$ is an integer, then one solution is

$$y_1 = x^{r_1} \sum_{k=0}^{\infty} a_k x^k,$$

and the second linearly independent solution is of the form

$$y_2 = x^{r_2} \sum_{k=0}^{\infty} b_k x^k + C(\ln x)y_1,$$

where we plug y_2 into (7.3) and solve for the constants b_k and C .

- (vi) Finally, if the indicial equation has complex roots, then solving for a_k in the solution

$$y = x^{r_1} \sum_{k=0}^{\infty} a_k x^k$$

results in a complex-valued function—all the a_k are complex numbers. We obtain our two linearly independent solutions* by taking the real and imaginary parts of y .

The main idea is to find at least one Frobenius-type solution. If we are lucky and find two, we are done. If we only get one, we either use the ideas above or even a different method such as reduction of order (see § 2.1) to obtain a second solution.

*See Joseph L. Neuringera, *The Frobenius method for complex roots of the indicial equation*, International Journal of Mathematical Education in Science and Technology, Volume 9, Issue 1, 1978, 71–77.

7.3.3 Bessel functions

An important class of functions that arises commonly in physics are the *Bessel functions*^{*}. For example, these functions appear when solving the wave equation in two and three dimensions. First consider *Bessel's equation* of order p :

$$x^2y'' + xy' + (x^2 - p^2)y = 0.$$

We allow p to be any number, not just an integer, although integers and multiples of $1/2$ are most important in applications.

When we plug

$$y = \sum_{k=0}^{\infty} a_k x^{k+r}$$

into Bessel's equation of order p , we obtain the indicial equation

$$r(r-1) + r - p^2 = (r-p)(r+p) = 0.$$

Therefore we obtain two roots $r_1 = p$ and $r_2 = -p$. If p is not an integer, then following the method of Frobenius and setting $a_0 = 1$, we obtain linearly independent solutions of the form

$$\begin{aligned} y_1 &= x^p \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{2^{2k} k! (k+p)(k-1+p) \cdots (2+p)(1+p)}, \\ y_2 &= x^{-p} \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{2^{2k} k! (k-p)(k-1-p) \cdots (2-p)(1-p)}. \end{aligned}$$

Exercise 7.3.1:

- a) Verify that the indicial equation of Bessel's equation of order p is $(r-p)(r+p) = 0$.
- b) Suppose p is not an integer. Carry out the computation to obtain the solutions y_1 and y_2 above.

Bessel functions are convenient constant multiples of y_1 and y_2 . First we must define the *gamma function*

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.$$

Notice that $\Gamma(1) = 1$. The gamma function also has a wonderful property

$$\Gamma(x+1) = x\Gamma(x).$$

From this property, it follows that $\Gamma(n) = (n-1)!$ when n is an integer. So the gamma function is a continuous version of the factorial. We compute:

$$\begin{aligned} \Gamma(k+p+1) &= (k+p)(k-1+p) \cdots (2+p)(1+p)\Gamma(1+p), \\ \Gamma(k-p+1) &= (k-p)(k-1-p) \cdots (2-p)(1-p)\Gamma(1-p). \end{aligned}$$

*Named after the German astronomer and mathematician [Friedrich Wilhelm Bessel](#) (1784–1846).

Exercise 7.3.2: Verify the identities above using $\Gamma(x+1) = x\Gamma(x)$.

We define the *Bessel functions of the first kind* of order p and $-p$ as

$$\begin{aligned} J_p(x) &= \frac{1}{2^p \Gamma(1+p)} y_1 = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(k+p+1)} \left(\frac{x}{2}\right)^{2k+p}, \\ J_{-p}(x) &= \frac{1}{2^{-p} \Gamma(1-p)} y_2 = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(k-p+1)} \left(\frac{x}{2}\right)^{2k-p}. \end{aligned}$$

As these are constant multiples of the solutions we found above, these are both solutions to Bessel's equation of order p . The constants are picked for convenience.

When p is not an integer, J_p and J_{-p} are linearly independent. When n is an integer we obtain

$$J_n(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! (k+n)!} \left(\frac{x}{2}\right)^{2k+n}.$$

In this case

$$J_n(x) = (-1)^n J_{-n}(x),$$

and so J_{-n} is not a second linearly independent solution. The other solution is the so-called *Bessel function of second kind*. These make sense only for integer orders n and are defined as limits of linear combinations of $J_p(x)$ and $J_{-p}(x)$, as p approaches n in the following way:

$$Y_n(x) = \lim_{p \rightarrow n} \frac{\cos(p\pi) J_p(x) - J_{-p}(x)}{\sin(p\pi)}.$$

Each linear combination of $J_p(x)$ and $J_{-p}(x)$ is a solution to Bessel's equation of order p . Then as we take the limit as p goes to n , we see that $Y_n(x)$ is a solution to Bessel's equation of order n . It also turns out that $Y_n(x)$ and $J_n(x)$ are linearly independent. Therefore when n is an integer, we have the general solution to Bessel's equation of order n :

$$y = AJ_n(x) + BY_n(x),$$

for arbitrary constants A and B . Note that $Y_n(x)$ goes to negative infinity at $x = 0$. Many mathematical software packages have these functions $J_n(x)$ and $Y_n(x)$ defined, so they can be used just like say $\sin(x)$ and $\cos(x)$. In fact, Bessel functions have some similar properties. For example, $-J_1(x)$ is a derivative of $J_0(x)$, and in general the derivative of $J_n(x)$ can be written as a linear combination of $J_{n-1}(x)$ and $J_{n+1}(x)$. Furthermore, these functions oscillate, although they are not periodic. See Figure 7.4 on the facing page for graphs of Bessel functions.

Example 7.3.4: Other equations can sometimes be solved in terms of the Bessel functions. For example, given a positive constant λ ,

$$xy'' + y' + \lambda^2 xy = 0,$$

can be changed to $x^2 y'' + xy' + \lambda^2 x^2 y = 0$. Then changing variables $t = \lambda x$, we obtain via chain rule the equation in y and t :

$$t^2 y'' + ty' + t^2 y = 0,$$

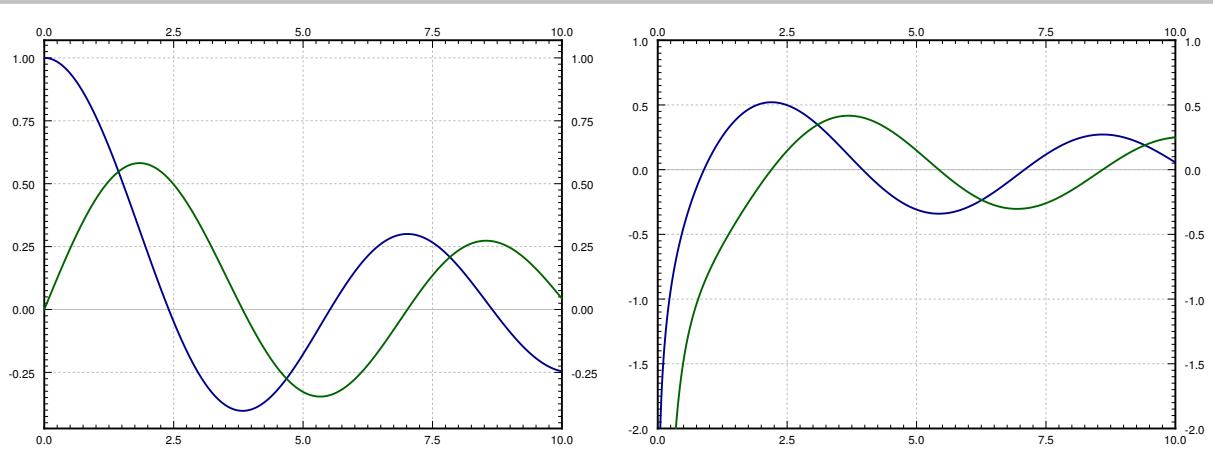


Figure 7.4: Plot of the $J_0(x)$ and $J_1(x)$ in the first graph and $Y_0(x)$ and $Y_1(x)$ in the second graph.

which we recognize as Bessel's equation of order 0. Therefore the general solution is $y(t) = AJ_0(t) + BY_0(t)$, or in terms of x :

$$y = AJ_0(\lambda x) + BY_0(\lambda x).$$

This equation comes up, for example, when finding the fundamental modes of vibration of a circular drum, but we digress.

7.3.4 Exercises

Exercise 7.3.3: Find a particular (Frobenius-type) solution of $x^2y'' + xy' + (1+x)y = 0$.

Exercise 7.3.4: Find a particular (Frobenius-type) solution of $xy'' - y = 0$.

Exercise 7.3.5:* Find a particular solution of $x^2y'' + (x - 3/4)y = 0$.

Exercise 7.3.6: Find a particular (Frobenius-type) solution of $y'' + \frac{1}{x}y' - xy = 0$.

Exercise 7.3.7: Find the general solution of $2xy'' + y' - x^2y = 0$.

Exercise 7.3.8:* Find the general solution of $x^2y'' - y = 0$.

Exercise 7.3.9: Find the general solution of $x^2y'' - xy' - y = 0$.

Exercise 7.3.10 (tricky):* Find the general solution of $x^2y'' - xy' + y = 0$.

Exercise 7.3.11: In the following equations classify the point $x = 0$ as ordinary, regular singular, or singular but not regular singular.

a) $x^2(1+x^2)y'' + xy = 0$

b) $x^2y'' + y' + y = 0$

c) $xy'' + x^3y' + y = 0$

d) $xy'' + xy' - e^x y = 0$

e) $x^2y'' + x^2y' + x^2y = 0$

Exercise 7.3.12:* In the following equations classify the point $x = 0$ as ordinary, regular singular, or singular but not regular singular.

a) $y'' + y = 0$

b) $x^3y'' + (1+x)y = 0$

c) $xy'' + x^5y' + y = 0$

d) $\sin(x)y'' - y = 0$

e) $\cos(x)y'' - \sin(x)y = 0$

Chapter 8

Inner Products and Orthogonality

8.1 Inner product and projections

Attribution: [JL], §A.5.

Learning Objectives

After this section, you will be able to:

- Compute the inner product of two vectors,
- Use the inner product to determine the angle between two vectors,
- Determine if two vectors are orthogonal using the inner product, and
- Use the Gram-Schmidt method to find an orthogonal basis for a given subspace.

8.1.1 Inner product and orthogonality

To do basic geometry, we need length, and we need angles. We have already seen the euclidean length, so let us figure out how to compute angles. Mostly, we are worried about the right angle[†].

Given two (column) vectors in \mathbb{R}^n , we define the (standard) *inner product* as the dot product:

$$\langle \vec{x}, \vec{y} \rangle = \vec{x} \cdot \vec{y} = \vec{y}^T \vec{x} = \sum_{i=1}^n x_i y_i.$$

Why do we seemingly give a new notation for the dot product? Because there are other possible inner products, which are not the dot product, although we will not worry about others here. An inner product can even be defined on spaces of functions as we do in chapter 9:

$$\langle f(t), g(t) \rangle = \int_a^b f(t)g(t) dt.$$

But we digress.

[†]When Euclid defined angles in his *Elements*, the only angle he ever really defined was the right angle.

Inner product satisfies the following rules

- (i) $\langle \vec{x}, \vec{x} \rangle \geq 0$, and $\langle \vec{x}, \vec{x} \rangle = 0$ if and only if $\vec{x} = 0$,
- (ii) $\langle \vec{x}, \vec{y} \rangle = \langle \vec{y}, \vec{x} \rangle$,
- (iii) $\langle a\vec{x}, \vec{y} \rangle = \langle \vec{x}, a\vec{y} \rangle = a\langle \vec{x}, \vec{y} \rangle$,
- (iv) $\langle \vec{x} + \vec{y}, \vec{z} \rangle = \langle \vec{x}, \vec{z} \rangle + \langle \vec{y}, \vec{z} \rangle$ and $\langle \vec{x}, \vec{y} + \vec{z} \rangle = \langle \vec{x}, \vec{y} \rangle + \langle \vec{x}, \vec{z} \rangle$.

In fact, anything that satisfies the properties above can be called an inner product, although in this section we are concerned with the standard inner product in \mathbb{R}^n .

The standard inner product gives the euclidean length:

$$\|\vec{x}\| = \sqrt{\langle \vec{x}, \vec{x} \rangle} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

How does it give angles?

You may recall from multivariable calculus, that in two or three dimensions, the standard inner product gives you the angle as you know from plane or three dimensional geometry:

$$\langle \vec{x}, \vec{y} \rangle = \|\vec{x}\| \|\vec{y}\| \cos \theta.$$

That is, θ is the angle that \vec{x} and \vec{y} make when they are based at the same point.

In \mathbb{R}^n , we are simply going to say that θ from the formula is what the angle is. This makes sense as any two vectors based at the origin lie in a 2-dimensional plane (subspace), and the formula works in 2 dimensions. In fact, one could even talk about angles between functions this way, and one can even discuss things like orthogonal functions.

To compute the angle we compute

$$\cos \theta = \frac{\langle \vec{x}, \vec{y} \rangle}{\|\vec{x}\| \|\vec{y}\|}.$$

Our angles are always in radians. We are computing the cosine of the angle, which is really the best we can do. Given two vectors at an angle θ , we can give the angle as $-\theta$, $2\pi - \theta$, etc., see [Figure 8.1](#) on the next page. Fortunately, $\cos \theta = \cos(-\theta) = \cos(2\pi - \theta)$. If we solve for θ using the inverse cosine \cos^{-1} , we can just decree that $0 \leq \theta \leq \pi$.

Example 8.1.1: Let us compute the angle between the vectors $(3, 0)$ and $(1, 1)$ in the plane. Compute

$$\cos \theta = \frac{\langle (3, 0), (1, 1) \rangle}{\|(3, 0)\| \|(1, 1)\|} = \frac{3+0}{3\sqrt{2}} = \frac{1}{\sqrt{2}}.$$

Therefore $\theta = \pi/4$.

As we said, the most important angle is the right angle. A right angle is $\pi/2$ radians, and $\cos(\pi/2) = 0$, so the formula is particularly easy in this case. We say vectors \vec{x} and \vec{y} are *orthogonal* if they are at right angles, that is if

$$\langle \vec{x}, \vec{y} \rangle = 0.$$

The vectors $(1, 0, 0, 1)$ and $(1, 2, 3, -1)$ are orthogonal. So are $(1, 1)$ and $(1, -1)$. However, $(1, 1)$ and $(1, 2)$ are not orthogonal as their inner product is 3 and not 0.

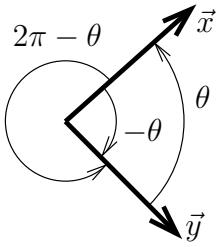


Figure 8.1: Angle between vectors.

8.1.2 Orthogonal projection

A typical application of linear algebra is to take a difficult problem, write everything in the right basis, and in this new basis the problem becomes simple. A particularly useful basis is an orthogonal basis, that is a basis where all the basis vectors are orthogonal. When we draw a coordinate system in two or three dimensions, we almost always draw our axes as orthogonal to each other.

Generalizing this concept to functions, it is particularly useful in [chapter 9](#) to express a function using a particular orthogonal basis, the Fourier series.

To express one vector in terms of an orthogonal basis, we need to first *project* one vector onto another. Given a nonzero vector \vec{v} , we define the *orthogonal projection* of \vec{w} onto \vec{v} as

$$\text{proj}_{\vec{v}}(\vec{w}) = \left(\frac{\langle \vec{w}, \vec{v} \rangle}{\langle \vec{v}, \vec{v} \rangle} \right) \vec{v}.$$

For the geometric idea, see [Figure 8.2](#). That is, we find the “shadow of \vec{w} ” on the line spanned by \vec{v} if the direction of the sun’s rays were exactly perpendicular to the line. Another way of thinking about it is that the tip of the arrow of $\text{proj}_{\vec{v}}(\vec{w})$ is the closest point on the line spanned by \vec{v} to the tip of the arrow of \vec{w} . In terms of Euclidean distance, $\vec{u} = \text{proj}_{\vec{v}}(\vec{w})$ minimizes the distance $\|\vec{w} - \vec{u}\|$ among all vectors \vec{u} that are multiples of \vec{v} . Because of this, this projection comes up often in applied mathematics in all sorts of contexts we cannot solve a problem exactly: We can’t always solve “Find \vec{w} as a multiple of \vec{v} ,” but $\text{proj}_{\vec{v}}(\vec{w})$ is the best “solution.”

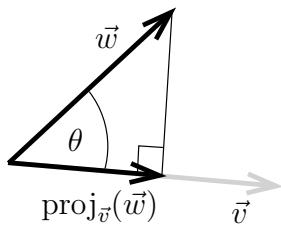


Figure 8.2: Orthogonal projection.

The formula follows from basic trigonometry. The length of $\text{proj}_{\vec{v}}(\vec{w})$ should be $\cos \theta$ times the length of \vec{w} , that is $(\cos \theta) \|\vec{w}\|$. We take the unit vector in the direction of \vec{v} , that

is, $\frac{\vec{v}}{\|\vec{v}\|}$ and we multiply it by the length of the projection. In other words,

$$\text{proj}_{\vec{v}}(\vec{w}) = (\cos \theta) \|\vec{w}\| \frac{\vec{v}}{\|\vec{v}\|} = \frac{(\cos \theta) \|\vec{w}\| \|\vec{v}\|}{\|\vec{v}\|^2} \vec{v} = \frac{\langle \vec{w}, \vec{v} \rangle}{\langle \vec{v}, \vec{v} \rangle} \vec{v}.$$

Example 8.1.2: Suppose we wish to project the vector $(1, 2, 3)$ onto the vector $(3, 2, 1)$. Compute

$$\begin{aligned} \text{proj}_{(1,2,3)}((3, 2, 1)) &= \frac{\langle (3, 2, 1), (1, 2, 3) \rangle}{\langle (1, 2, 3), (1, 2, 3) \rangle} (1, 2, 3) = \frac{3 \cdot 1 + 2 \cdot 2 + 1 \cdot 3}{1 \cdot 1 + 2 \cdot 2 + 3 \cdot 3} (1, 2, 3) \\ &= \frac{10}{14} (1, 2, 3) = \left(\frac{5}{7}, \frac{10}{7}, \frac{15}{7} \right). \end{aligned}$$

Let us double check that the projection is orthogonal. That is $\vec{w} - \text{proj}_{\vec{v}}(\vec{w})$ ought to be orthogonal to \vec{v} , see the right angle in [Figure 8.2](#) on the previous page. That is,

$$(3, 2, 1) - \text{proj}_{(1,2,3)}((3, 2, 1)) = \left(3 - \frac{5}{7}, 2 - \frac{10}{7}, 1 - \frac{15}{7} \right) = \left(\frac{16}{7}, \frac{4}{7}, \frac{-8}{7} \right)$$

ought to be orthogonal to $(1, 2, 3)$. We compute the inner product and we had better get zero:

$$\left\langle \left(\frac{16}{7}, \frac{4}{7}, \frac{-8}{7} \right), (1, 2, 3) \right\rangle = \frac{16}{7} \cdot 1 + \frac{4}{7} \cdot 2 - \frac{8}{7} \cdot 3 = 0.$$

8.1.3 Orthogonal basis

As we said, a basis $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$ is an *orthogonal basis* if all vectors in the basis are orthogonal to each other, that is, if

$$\langle \vec{v}_j, \vec{v}_k \rangle = 0$$

for all choices of j and k where $j \neq k$ (a nonzero vector cannot be orthogonal to itself). A basis is furthermore called an *orthonormal basis* if all the vectors in a basis are also unit vectors, that is, if all the vectors have magnitude 1. For example, the standard basis $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ is an orthonormal basis of \mathbb{R}^3 : Any pair is orthogonal, and each vector is of unit magnitude.

The reason why we are interested in orthogonal (or orthonormal) bases is that they make it really simple to represent a vector (or a projection onto a subspace) in the basis. The simple formula for the orthogonal projection onto a vector gives us the coefficients. In [chapter 9](#) we use the same idea by finding the correct orthogonal basis for the set of solutions of a differential equation we are then able to find any particular solution by simply applying the orthogonal projection formula, which is just a couple of inner products.

Let us come back to linear algebra. Suppose that we have a subspace and an orthogonal basis $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$. We wish to express \vec{x} in terms of the basis. If \vec{x} is not in the span of the basis (when it is not in the given subspace), then of course it is not possible, but the following formula gives us at least the orthogonal projection onto the subspace.

First suppose that \vec{x} is in the span. Then it is the sum of the orthogonal projections:

$$\vec{x} = \text{proj}_{\vec{v}_1}(\vec{x}) + \text{proj}_{\vec{v}_2}(\vec{x}) + \cdots + \text{proj}_{\vec{v}_n}(\vec{x}) = \frac{\langle \vec{x}, \vec{v}_1 \rangle}{\langle \vec{v}_1, \vec{v}_1 \rangle} \vec{v}_1 + \frac{\langle \vec{x}, \vec{v}_2 \rangle}{\langle \vec{v}_2, \vec{v}_2 \rangle} \vec{v}_2 + \cdots + \frac{\langle \vec{x}, \vec{v}_n \rangle}{\langle \vec{v}_n, \vec{v}_n \rangle} \vec{v}_n.$$

In other words, if we want to write $\vec{x} = a_1 \vec{v}_1 + a_2 \vec{v}_2 + \cdots + a_n \vec{v}_n$, then

$$a_1 = \frac{\langle \vec{x}, \vec{v}_1 \rangle}{\langle \vec{v}_1, \vec{v}_1 \rangle}, \quad a_2 = \frac{\langle \vec{x}, \vec{v}_2 \rangle}{\langle \vec{v}_2, \vec{v}_2 \rangle}, \quad \dots, \quad a_n = \frac{\langle \vec{x}, \vec{v}_n \rangle}{\langle \vec{v}_n, \vec{v}_n \rangle}.$$

Another way to derive this formula is to work in reverse. Suppose that $\vec{x} = a_1 \vec{v}_1 + a_2 \vec{v}_2 + \cdots + a_n \vec{v}_n$. Take an inner product with \vec{v}_j , and use the properties of the inner product:

$$\begin{aligned} \langle \vec{x}, \vec{v}_j \rangle &= \langle a_1 \vec{v}_1 + a_2 \vec{v}_2 + \cdots + a_n \vec{v}_n, \vec{v}_j \rangle \\ &= a_1 \langle \vec{v}_1, \vec{v}_j \rangle + a_2 \langle \vec{v}_2, \vec{v}_j \rangle + \cdots + a_n \langle \vec{v}_n, \vec{v}_j \rangle. \end{aligned}$$

As the basis is orthogonal, then $\langle \vec{v}_k, \vec{v}_j \rangle = 0$ whenever $k \neq j$. That means that only one of the terms, the j^{th} one, on the right hand side is nonzero and we get

$$\langle \vec{x}, \vec{v}_j \rangle = a_j \langle \vec{v}_j, \vec{v}_j \rangle.$$

Solving for a_j we find $a_j = \frac{\langle \vec{x}, \vec{v}_j \rangle}{\langle \vec{v}_j, \vec{v}_j \rangle}$ as before.

Example 8.1.3: The vectors $(1, 1)$ and $(1, -1)$ form an orthogonal basis of \mathbb{R}^2 . Find the coefficients that represent $(3, 4)$ in terms of this basis, that is, we wish to find a_1 and a_2 such that

$$(3, 4) = a_1(1, 1) + a_2(1, -1).$$

Solution: We compute:

$$a_1 = \frac{\langle (3, 4), (1, 1) \rangle}{\langle (1, 1), (1, 1) \rangle} = \frac{7}{2}, \quad a_2 = \frac{\langle (3, 4), (1, -1) \rangle}{\langle (1, -1), (1, -1) \rangle} = \frac{-1}{2}.$$

So

$$(3, 4) = \frac{7}{2}(1, 1) + \frac{-1}{2}(1, -1). \quad \square$$

If the basis is orthonormal rather than orthogonal, than the denominators are always just one. It is easy to make a basis orthonormal, just by dividing all the vectors by their size. If you want to decompose many vectors, it may be better to find an orthonormal basis. In the example above, the orthonormal basis we would thus create is

$$\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right), \quad \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}} \right).$$

Then the computation would have been

$$\begin{aligned} (3, 4) &= \left\langle (3, 4), \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) \right\rangle \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) + \left\langle (3, 4), \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}} \right) \right\rangle \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}} \right) \\ &= \frac{7}{\sqrt{2}} \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right) + \frac{-1}{\sqrt{2}} \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}} \right). \end{aligned}$$

Maybe the example is not so awe inspiring, but given vectors in \mathbb{R}^{20} rather than \mathbb{R}^2 , then surely one would much rather do 20 inner products (or 40 if we did not have an orthonormal basis) rather than solving a system of twenty equations in twenty unknowns using row reduction of a 20×21 matrix.

As we said above, the formula still works even if \vec{x} is not in the subspace, although then it does not get us the vector \vec{x} but its projection. More concretely, suppose that S is a subspace that is the span of $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$ and \vec{x} is any vector. Let $\text{proj}_S(\vec{x})$ be the vector in S that is the closest to \vec{x} . Then

$$\text{proj}_S(\vec{x}) = \frac{\langle \vec{x}, \vec{v}_1 \rangle}{\langle \vec{v}_1, \vec{v}_1 \rangle} \vec{v}_1 + \frac{\langle \vec{x}, \vec{v}_2 \rangle}{\langle \vec{v}_2, \vec{v}_2 \rangle} \vec{v}_2 + \cdots + \frac{\langle \vec{x}, \vec{v}_n \rangle}{\langle \vec{v}_n, \vec{v}_n \rangle} \vec{v}_n.$$

Of course, if \vec{x} is in S , then $\text{proj}_S(\vec{x}) = \vec{x}$, as the closest vector in S to \vec{x} is \vec{x} itself. But true utility is obtained when \vec{x} is not in S . In much of applied mathematics we cannot find an exact solution to a problem, but we try to find the best solution out of a small subset (subspace). The partial sums of Fourier series are one example. Another example is least square approximation to fit a curve to data. Yet another example is given by the most commonly used numerical methods to solve differential equations, the finite element methods.

Example 8.1.4: The vectors $(1, 2, 3)$ and $(3, 0, -1)$ are orthogonal, and so they are an orthogonal basis of a subspace S :

$$S = \text{span}\{(1, 2, 3), (3, 0, -1)\}.$$

Let us find the vector in S that is closest to $(2, 1, 0)$. That is, let us find $\text{proj}_S((2, 1, 0))$.

$$\begin{aligned} \text{proj}_S((2, 1, 0)) &= \frac{\langle (2, 1, 0), (1, 2, 3) \rangle}{\langle (1, 2, 3), (1, 2, 3) \rangle} (1, 2, 3) + \frac{\langle (2, 1, 0), (3, 0, -1) \rangle}{\langle (3, 0, -1), (3, 0, -1) \rangle} (3, 0, -1) \\ &= \frac{2}{7}(1, 2, 3) + \frac{3}{5}(3, 0, -1) \\ &= \left(\frac{73}{35}, \frac{4}{7}, \frac{9}{35} \right). \end{aligned}$$

8.1.4 The Gram–Schmidt process

Before leaving orthogonal bases, let us note a procedure for manufacturing them out of any old basis. It may not be difficult to come up with an orthogonal basis for a 2-dimensional subspace, but for a 20-dimensional subspace, it seems a daunting task. Fortunately, the orthogonal projection can be used to “project away” the bits of the vectors that are making them not orthogonal. It is called the *Gram–Schmidt process*.

We start with a basis of vectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$. We construct an orthogonal basis $\vec{w}_1, \vec{w}_2, \dots, \vec{w}_n$ as follows.

$$\begin{aligned} \vec{w}_1 &= \vec{v}_1, \\ \vec{w}_2 &= \vec{v}_2 - \text{proj}_{\vec{w}_1}(\vec{v}_2), \\ \vec{w}_3 &= \vec{v}_3 - \text{proj}_{\vec{w}_1}(\vec{v}_3) - \text{proj}_{\vec{w}_2}(\vec{v}_3), \end{aligned}$$

$$\begin{aligned}\vec{w}_4 &= \vec{v}_4 - \text{proj}_{\vec{w}_1}(\vec{v}_4) - \text{proj}_{\vec{w}_2}(\vec{v}_4) - \text{proj}_{\vec{w}_3}(\vec{v}_4), \\ &\vdots \\ \vec{w}_n &= \vec{v}_n - \text{proj}_{\vec{w}_1}(\vec{v}_n) - \text{proj}_{\vec{w}_2}(\vec{v}_n) - \cdots - \text{proj}_{\vec{w}_{n-1}}(\vec{v}_n).\end{aligned}$$

What we do is at the k^{th} step, we take \vec{v}_k and we subtract the projection of \vec{v}_k to the subspace spanned by $\vec{w}_1, \vec{w}_2, \dots, \vec{w}_{k-1}$.

Example 8.1.5: Consider the vectors $(1, 2, -1)$, and $(0, 5, -2)$ and call S the span of the two vectors. Let us find an orthogonal basis of S :

$$\begin{aligned}\vec{w}_1 &= (1, 2, -1), \\ \vec{w}_2 &= (0, 5, -2) - \text{proj}_{(1, 2, -1)}((0, 2, -2)) \\ &= (0, 1, -1) - \frac{\langle (0, 5, -2), (1, 2, -1) \rangle}{\langle (1, 2, -1), (1, 2, -1) \rangle}(1, 2, -1) = (0, 5, -2) - 2(1, 2, -1) = (-2, 1, 0).\end{aligned}$$

So $(1, 2, -1)$ and $(-2, 1, 0)$ span S and are orthogonal. Let us check: $(1, 2, -1) \cdot (-2, 1, 0) = 0$.

Suppose we wish to find an orthonormal basis, not just an orthogonal one. Well, we simply make the vectors into unit vectors by dividing them by their magnitude. The two vectors making up the orthonormal basis of S are:

$$\frac{1}{\sqrt{6}}(1, 2, -1) = \left(\frac{1}{\sqrt{6}}, \frac{2}{\sqrt{6}}, \frac{-1}{\sqrt{6}} \right), \quad \frac{1}{\sqrt{5}}(-2, 1, 0) = \left(\frac{-2}{\sqrt{5}}, \frac{1}{\sqrt{5}}, 0 \right).$$

8.1.5 Exercises

Exercise 8.1.1: Find the s that makes the following vectors orthogonal: $(1, 2, 3)$, $(1, 1, s)$.

Exercise 8.1.2:* Find the s that makes the following vectors orthogonal: $(1, 1, 1)$, $(1, s, 1)$.

Exercise 8.1.3: Find the angle θ between $(1, 3, 1)$, $(2, 1, -1)$.

Exercise 8.1.4:* Find the angle θ between $(1, 2, 3)$, $(1, 1, 1)$.

Exercise 8.1.5: Given that $\langle \vec{v}, \vec{w} \rangle = 3$ and $\langle \vec{v}, \vec{u} \rangle = -1$ compute

$$\text{a) } \langle \vec{u}, 2\vec{v} \rangle \quad \text{b) } \langle \vec{v}, 2\vec{w} + 3\vec{u} \rangle \quad \text{c) } \langle \vec{w} + 3\vec{u}, \vec{v} \rangle$$

Exercise 8.1.6:* Given that $\langle \vec{v}, \vec{w} \rangle = 1$ and $\langle \vec{v}, \vec{u} \rangle = -1$ and $\|\vec{v}\| = 3$ and

$$\text{a) } \langle 3\vec{u}, 5\vec{v} \rangle \quad \text{b) } \langle \vec{v}, 2\vec{w} + 3\vec{u} \rangle \quad \text{c) } \langle \vec{w} + 3\vec{v}, \vec{v} \rangle$$

Exercise 8.1.7: Suppose $\vec{v} = (1, 1, -1)$. Find

$$\text{a) } \text{proj}_{\vec{v}}((1, 0, 0)) \quad \text{b) } \text{proj}_{\vec{v}}((1, 2, 3)) \quad \text{c) } \text{proj}_{\vec{v}}((1, -1, 0))$$

Exercise 8.1.8:* Suppose $\vec{v} = (1, 0, -1)$. Find

$$\text{a) } \text{proj}_{\vec{v}}((0, 2, 1)) \quad \text{b) } \text{proj}_{\vec{v}}((1, 0, 1)) \quad \text{c) } \text{proj}_{\vec{v}}((4, -1, 0))$$

Exercise 8.1.9: Consider the vectors $(1, 2, 3)$, $(-3, 0, 1)$, $(1, -5, 3)$.

- a) Check that the vectors are linearly independent and so form a basis.
- b) Check that the vectors are mutually orthogonal, and are therefore an orthogonal basis.
- c) Represent $(1, 1, 1)$ as a linear combination of this basis.
- d) Make the basis orthonormal.

Exercise 8.1.10:* The vectors $(1, 1, -1)$, $(2, -1, 1)$, $(1, -5, 3)$ form an orthonormal basis. Represent the following vectors in terms of this basis.

- a) $(1, -8, 4)$
- b) $(5, -7, 5)$
- c) $(0, -6, 2)$

Exercise 8.1.11: Let S be the subspace spanned by $(1, 3, -1)$, $(1, 1, 1)$. Find an orthogonal basis of S by the Gram-Schmidt process.

Exercise 8.1.12:* Let S be the subspace spanned by $(2, -1, 1)$, $(2, 2, 2)$. Find an orthogonal basis of S by the Gram-Schmidt process.

Exercise 8.1.13: Starting with $(1, 2, 3)$, $(1, 1, 1)$, $(2, 2, 0)$, follow the Gram-Schmidt process to find an orthogonal basis of \mathbb{R}^3 .

Exercise 8.1.14:* Starting with $(1, 1, -1)$, $(2, 3, -1)$, $(1, -1, 1)$, follow the Gram-Schmidt process to find an orthogonal basis of \mathbb{R}^3 .

Exercise 8.1.15: Find an orthogonal basis of \mathbb{R}^3 such that $(3, 1, -2)$ is one of the vectors. Hint: First find two extra vectors to make a linearly independent set.

Exercise 8.1.16: Using cosines and sines of θ , find a unit vector \vec{u} in \mathbb{R}^2 that makes angle θ with $\vec{v} = (1, 0)$. What is $\langle \vec{v}, \vec{u} \rangle$?

8.2 Special Matrices

Learning Objectives

After this section, you will be able to:

- Compute the adjoint of a matrix,
- Determine if a matrix is orthogonal, and
- Diagonalize a matrix using eigenvalues and eigenvectors if possible.

In this section, we will discuss several different types of matrices that have nice properties related to the inner product on vectors.

8.2.1 Adjoint Matrix

To begin this section, we want to discuss how matrices, when applied to vectors, interact with the inner product on those vectors.

Definition 8.2.1

Let A be an $n \times n$ matrix. The *adjoint* matrix of A , written A^* , is another $n \times n$ matrix so that for all n -component vectors \vec{v} and \vec{w}

$$\langle A\vec{v}, \vec{w} \rangle = \langle \vec{v}, A^*\vec{w} \rangle.$$

The idea of the adjoint is that if we want to move a matrix to the other side of an inner product, we need to convert it to the adjoint.

Example 8.2.1: Compute the adjoint of the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 3 & -1 \end{bmatrix}.$$

Solution: All we have to go from is the definition of the adjoint. Let's start with two vectors

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad \vec{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

and as a place-holder, set

$$A^* = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

because we know it is a 2×2 matrix. We want to satisfy

$$\langle A\vec{v}, \vec{w} \rangle = \langle \vec{v}, A^*\vec{w} \rangle.$$

With this, we can compute that

$$A\vec{v} = \begin{bmatrix} 1v_1 + 2v_2 \\ 3v_1 - v_2 \end{bmatrix}$$

so that

$$\langle A\vec{v}, \vec{w} \rangle = \left\langle \begin{bmatrix} 1v_1 + 2v_2 \\ 3v_1 - v_2 \end{bmatrix}, \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \right\rangle = v_1w_1 + 2v_2w_1 + 3v_1w_2 - v_2w_2. \quad (8.1)$$

Similarly, we have that

$$A^*\vec{w} = \begin{bmatrix} aw_1 + bw_2 \\ cw_1 + dw_2 \end{bmatrix}$$

so that

$$\langle \vec{v}, A^*\vec{w} \rangle = \left\langle \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}, \begin{bmatrix} aw_1 + bw_2 \\ cw_1 + dw_2 \end{bmatrix} \right\rangle = aw_1v_1 + bv_1w_2 + cw_1v_1 + dv_2w_2. \quad (8.2)$$

In order to make (8.1) and (8.2) match, we need to set $a = 1$, $b = 3$, $c = 2$ and $d = -1$. Thus, the adjoint matrix A^* is

$$A^* = \begin{bmatrix} 1 & 3 \\ 2 & -1 \end{bmatrix}. \quad \boxed{}$$

This result looks fairly similar to the original matrix A ; in fact, it looks like it is the transpose of the matrix A . It turns out that this is always the case.

Theorem 8.2.1

Let A be a *real* $n \times n$ matrix, and $\langle \cdot, \cdot \rangle$ the standard inner product on n component real vectors. Then $A^* = A^T$.

Proof. The idea is exactly the same as the example that we went through. If we take real vectors \vec{v} and \vec{w} with components v_i and w_i respectively, and let A be the $n \times n$ matrix with components a_{ij} , we can write out the inner product as a sum.

$$\begin{aligned} \langle A\vec{v}, \vec{w} \rangle &= \left\langle \sum_{j=1}^n a_{ij}v_j, \vec{w} \right\rangle \\ &= \sum_{i=1}^n \left(\sum_{j=1}^n a_{ij}v_j \right) w_i \\ &= \sum_{i=1}^n \sum_{j=1}^n a_{ij}v_j w_i \end{aligned}$$

In order to get the adjoint matrix, we need to group the terms to match with $\langle \text{vec } v, A^*\vec{w} \rangle$, which means we want the \vec{v} part to be last. We can then see that

$$\langle A\vec{v}, \vec{w} \rangle = \sum_{i=1}^n \sum_{j=1}^n a_{ij}v_j w_i = \sum_{j=1}^n \left(\sum_{i=1}^n a_{ij}w_i \right) v_j = \langle \vec{v}, \sum_{i=1}^n a_{ij}w_i \rangle.$$

Thus, we have that

$$A^*\vec{w} = \sum_{j=1}^n a_{ij}w_i.$$

This is *almost* the matrix-vector product $A\vec{w}$, but the i and j are backwards. If we swap the i and j , that is the same as swapping the rows and columns of a matrix, which gives rise to the transpose. Therefore, for the matrix $B = A^T$, we have that

$$B\vec{w} = \sum_{i=1}^n b_{ji} w_i = \sum_{i=1}^n a_{ij} w_i = A^* \vec{w}.$$

Thus, we have that $A^* = A^T$ □

Remark 8.2.1: The same ideas work for complex vector spaces and complex inner products. In that case, the inner product is defined by

$$\langle \vec{v}, \vec{w} \rangle = \sum_{k=1}^n v_k \overline{w_k}$$

and the adjoint is

$$A^* = A^H,$$

which is the Hermitian transpose of A . This is sometimes also called the conjugate transpose; for the matrix A we take the transpose, and then take the complex conjugate of each entry of the matrix. The proof follows all of the same steps.

Example 8.2.2: Determine the adjoint of the following matrices.

$$A = \begin{bmatrix} 1 & -3 \\ 2 & 5 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 3 & 0 \\ 3 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}.$$

Solution: Since we know that the adjoint is just the transpose, we can compute these fairly easily.

$$A^* = \begin{bmatrix} 1 & 2 \\ -3 & 5 \end{bmatrix}$$

and

$$B^* = \begin{bmatrix} 1 & 3 & 0 \\ 3 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}.$$
□

The matrix B is a special matrix because the adjoint is the same as the original matrix. This is called a *self-adjoint ! matrix* matrix. The language of the adjoint will allow us to be able to talk about some more special matrices and how they interact with the inner product.

8.2.2 Orthogonal Matrices

We know how to discuss if two vectors are orthogonal, namely, that their inner product is zero. In general, we say that any set of vectors $\{\vec{v}_1, \dots, \vec{v}_k\}$ is *orthogonal* if any pair of them have inner product zero. This is the same idea of the orthogonal basis discussed in § 8.1, but

without the requirement that it span the entire space. If we also have that the length of each vector in this set is 1, that means that the set is *orthonormal*.

We can extend this definition of orthogonal vectors to talk about matrices.

Definition 8.2.2

An $n \times n$ matrix A is *orthogonal* if the columns of A are orthonormal. That is, if $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$ are the columns of A , then we have

- $\langle v_i, v_i \rangle = 1$ for all i ,
- $\langle v_i, v_j \rangle = 0$ for all $i \neq j$.

Note that for matrices, we are only talking about a single matrix being orthogonal, unlike for vectors where we talk about a pair or set of vectors being orthogonal.

Example 8.2.3: Determine if each of the following matrices are orthogonal.

$$A = \begin{bmatrix} \frac{3}{5} & 1 \\ \frac{4}{5} & 0 \\ \frac{5}{5} & 0 \end{bmatrix} \quad B = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Solution: For matrix A , we check the length of each column vector and then the inner product of the two vectors. For the first column we get that

$$\|\vec{v}_1\| = \sqrt{\left(\frac{3}{5}\right)^2 + \left(\frac{4}{5}\right)^2} = \sqrt{\frac{9}{25} + \frac{16}{25}} = 1$$

and $\|\vec{v}_2\| = 1$. So, these vectors are both length 1. However

$$\langle v_1, v_2 \rangle = \frac{3}{5}(1) + \frac{4}{5}(0) = \frac{3}{5} \neq 0,$$

so these vectors are not orthogonal. Therefore, A is not an orthogonal matrix.

For matrix B , each column has exactly one non-zero entry of absolute value 1, so the columns all have length 1. In addition, none of the columns have this non-zero entry in the same row, so if we were to take the inner product of any two columns, we would get zero. Therefore B is an orthogonal matrix. \square

So what are these orthogonal matrices? There is one main property of these matrices that drives their use and description. Let A be an orthogonal matrix and consider the matrix product $B = A^T A$. If we look at a single entry of B , the definition of matrix multiplication says that b_{ij} is the dot product of the i th row of A^T with the j th column of A . Since the first matrix is A^T , the i th row of A^T is the same as the i th column of A , so that b_{ij} is the dot product of the i th column of A with the j th column of A .

However, we know things about this. Since A is orthogonal, we know that the columns of A form an orthonormal set. Therefore, if $i \neq j$, the dot product is zero, and if they are the same, it is 1. Therefore, we know that $b_{ij} = 1$ if $i = j$ and 0 if $i \neq j$. Therefore, B is

the identity matrix. This means that the transpose of an orthogonal matrix is equal to its inverse.

Theorem 8.2.2

If A is an orthogonal matrix, then $A^T = A^{-1}$, so that

$$AA^T = A^TA = I.$$

By our discussion in the previous section, this means that the adjoint of an orthogonal matrix is also its inverse. This allows us to come up with several nice properties of orthogonal matrices.

Theorem 8.2.3

Let A be an $n \times n$ orthogonal matrix and \vec{v} and \vec{w} two n -component real vectors. Then

1. $\det(A) = \pm 1$
2. $\langle A\vec{v}, A\vec{w} \rangle = \langle \vec{v}, \vec{w} \rangle$
3. $\|A\vec{v}\| = \|\vec{v}\|$.

Proof. 1. Since we know that $AA^T = I$ and that $\det(A) = \det(A^T)$, we have that

$$1 = \det(I) = \det(AA^T) = \det(A)\det(A^T) = \det(A)^2.$$

Thus, we know that $\det(A)$ must be either 1 or -1 .

2. Using the properties of the adjoint, we know that

$$\langle A\vec{v}, A\vec{w} \rangle = \langle \vec{v}, A^*A\vec{w} \rangle.$$

However, $A^* = A^T = A^{-1}$, so that $A^*A = A^{-1}A = I$. Therefore, this expression equals $\langle \vec{v}, \vec{w} \rangle$, which is what we wanted.

3. Using the previous part, we have that

$$\|A\vec{v}\|^2 = \langle A\vec{v}, A\vec{v} \rangle = \langle \vec{v}, \vec{v} \rangle = \|\vec{v}\|^2,$$

which gives the desired result after taking square roots. □

What do these properties tell us? Since the inner product discusses angles between vectors, we see that orthogonal matrices are transformations that preserve lengths and angles. This means that they are composed of different “rigid motions”, either reflections or rotations around different axes. For the two dimensional case, we have the matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

which represents a reflection over the x -axis, and for any angle θ , the matrix

$$\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{bmatrix}$$

gives a rotation of angle θ in the counterclockwise direction.

Exercise 8.2.1: Verify that these two matrices are orthogonal.

These (and products of them) are the only options for 2×2 matrices. Once we get beyond that, there isn't necessarily a single rotation and reflection that can define everything, but the idea of an orthogonal matrix being composed of different rotations and reflections is still true.

8.2.3 Symmetric Matrices

Another type of matrix that interacts nicely with the inner product is a symmetric matrix. A matrix A is *symmetric* if $A = A^T$. For example, the matrix

$$A = \begin{bmatrix} 1 & 4 \\ 4 & 3 \end{bmatrix}$$

is symmetric, because if we swap the rows and columns, we get the original matrix back. However, the matrix

$$B = \begin{bmatrix} 1 & 2 & 3 \\ -1 & 2 & 1 \\ 0 & 3 & -2 \end{bmatrix}$$

is not symmetric, because

$$B^T = \begin{bmatrix} 1 & -1 & 0 \\ 2 & 2 & 3 \\ 3 & 1 & -2 \end{bmatrix},$$

which does not match B .

From our discussion of the adjoint, we know that for a real matrix A , the transpose is the same as the adjoint. This means that for symmetric matrices, the matrix also equals the adjoint, and is so all symmetric matrices are self-adjoint. This means that if we have a symmetric matrix A , then for any two vectors \vec{v} and \vec{w} ,

$$\langle A\vec{v}, \vec{w} \rangle = \langle \vec{v}, A\vec{w} \rangle.$$

So, if we have a symmetric matrix, we can move it to the other side of the inner product without changing anything. This has some nice consequences for these matrices, particularly with respect to eigenvalues.

Theorem 8.2.4

Let A be a symmetric real matrix. Let λ_1 and λ_2 be two *different* eigenvalues of A with corresponding eigenvectors \vec{v}_1 and \vec{v}_2 . Then \vec{v}_1 and \vec{v}_2 are orthogonal.

Proof. Consider the expression $\langle A\vec{v}_1, \vec{v}_2 \rangle$. We will compute this two ways. First is by using the fact that \vec{v}_1 is an eigenvector. This means that $A\vec{v}_1 = \lambda_1\vec{v}_1$ so

$$\langle A\vec{v}_1, \vec{v}_2 \rangle = \lambda_1 \langle \vec{v}_1, \vec{v}_2 \rangle. \quad (8.3)$$

For a second computation, we will use the fact the fact that A is self-adjoint and that \vec{v}_2 is an eigenvector. This gives

$$\langle A\vec{v}_1, \vec{v}_2 \rangle = \langle \vec{v}_1, A\vec{v}_2 \rangle = \langle \vec{v}_1, \lambda_2\vec{v}_2 \rangle = \lambda_2 \langle \vec{v}_1, \vec{v}_2 \rangle. \quad (8.4)$$

Since (8.3) and (8.4) must be equal, we have that

$$\lambda_1 \langle \vec{v}_1, \vec{v}_2 \rangle = \lambda_2 \langle \vec{v}_1, \vec{v}_2 \rangle$$

or, by subtracting to the other side,

$$(\lambda_1 - \lambda_2) \langle \vec{v}_1, \vec{v}_2 \rangle = 0.$$

Since $\lambda_1 \neq \lambda_2$, the only way this can be true is if $\langle \vec{v}_1, \vec{v}_2 \rangle = 0$. Therefore \vec{v}_1 and \vec{v}_2 are orthogonal. \square

Example 8.2.4: Find all of the eigenvalues and eigenvectors of the matrix

$$A = \begin{bmatrix} 3 & 0 \\ 0 & 5 & 2 \\ 0 & 2 & 2 \end{bmatrix}$$

and verify that the eigenvectors are orthogonal.

Solution: To find the eigenvalues of this matrix, we compute

$$\det(A - \lambda I) = (3 - \lambda)((5 - \lambda)(2 - \lambda) - 4) = (3 - \lambda)(\lambda^2 - 7\lambda + 6) = (3 - \lambda)(\lambda - 1)(\lambda - 6).$$

Therefore, the eigenvalues are 1, 3, and 6. For $\lambda = 1$, the system of equations $(A - I)\vec{v}_1 = 0$ becomes

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 2 \\ 0 & 2 & 1 \end{bmatrix} \vec{v} = 0$$

which has solution $\vec{v}_1 = \begin{bmatrix} 0 \\ -1 \\ 2 \end{bmatrix}$.

For $\lambda = 3$, the system becomes

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & -1 \end{bmatrix} \vec{v}_3 = 0$$

which has solution $\vec{v}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$.

Finally, for $\lambda = 6$, we get the system

$$\begin{bmatrix} -3 & 0 & 0 \\ 0 & -1 & 2 \\ 0 & 2 & -4 \end{bmatrix} \vec{v}_6 = 0$$

which has solution $\vec{v}_6 = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$.

We can then check that

$$\langle \vec{v}_1, \vec{v}_3 \rangle = 0 \quad \langle \vec{v}_1, \vec{v}_6 \rangle = 0 \quad \langle \vec{v}_3, \vec{v}_6 \rangle = 0$$

which tells us that the eigenvectors of this matrix are orthogonal. □

There are a few additional properties of symmetric matrices in terms of their eigenvalues and eigenvectors that we will not prove here. Firstly, the eigenvalues of a symmetric matrix are always real; that is, it is impossible to have a real symmetric matrix with complex eigenvalues. Secondly, a real symmetric matrix will always have a basis of eigenvectors. While this doesn't necessarily mean that the matrix does not have repeated eigenvalues, it means that none of these eigenvalues are defective, so the problems that come from them do not arise. This is particularly relevant for diagonalization, which we cover next.

8.2.4 Diagonalizable Matrices

We used the idea of diagonalization in § 4.6 to solve non-homogeneous systems of differential equations. However, the idea of being able to diagonalize a matrix has independent uses, so we analyze it further here.

Definition 8.2.3

A square matrix A is diagonalizable if there exists an invertible matrix P and a diagonal matrix D so that $A = PDP^{-1}$.

This essentially means that there is a set of coordinates in which A acts like a diagonal matrix, or a set of directions in which A acts like a scalar. We already know some directions where that happens, and those are the eigenvectors.

Theorem 8.2.5

Let A be an $n \times n$ square matrix. If there is a set of vectors $S = \{\vec{v}_1, \dots, \vec{v}_n\}$ so that

- S is a basis of \mathbb{R}^n
- Each of the vectors in S is an eigenvector of A ,

then A is diagonalizable.

Proof. Define the matrix P to be the matrix whose columns are the vectors $\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n$ in the set S and let D be the diagonal matrix whose entries are $\lambda_1, \lambda_2, \dots, \lambda_n$, which are the eigenvalues of each vector in S in the same order as those vectors. We want to show that $A = PDP^{-1}$. Firstly, we know that P is invertible since the vectors in S are linearly independent. Since the vectors in S form a basis for \mathbb{R}^n , we can do this by showing that $A\vec{v}_i = PDP^{-1}\vec{v}_i$ for each i from 1 to n . We know that, since \vec{v}_i is an eigenvector of A , we have that $A\vec{v}_i = \lambda_i\vec{v}_i$.

For the other side of the desired equality, we first look at $P^{-1}\vec{v}_i$. Assuming this equals the vector \vec{w} , then we have that $\vec{v}_i = P\vec{w}$. Since \vec{v}_i is the i th column of P , the properties of matrix multiplication means that we want \vec{w} to have a 1 in the i th component and 0 in every other component, which will extract the i th column of P . So, we have that $P^{-1}\vec{v}_i = \vec{e}_i$, where this denotes exactly the vector described previously.

The next step is to multiply by D . Since the i th column of D just has λ_i in the i th component, the product $D\vec{e}_i$ is $\lambda_i e_i$. Finally, we need to multiply this matrix by P , which will give us exactly λ_i times the i th column of P . Since that is exactly \vec{v}_i , the final product of $PDP^{-1}\vec{v}_i$ is $\lambda_i\vec{v}_i$, which is the same as we got from $A\vec{v}_i$.

Finally, to justify why this proves the two matrices are equal, we see that for any vector \vec{v} , there are coefficients c_1, \dots, c_n so that

$$\vec{v} = c_1\vec{v}_1 + c_2\vec{v}_2 + \cdots + c_n\vec{v}_n$$

because S is a basis of \mathbb{R}^n . Then, by linearity of matrix multiplication, we have that

$$\begin{aligned} A\vec{v} &= A(c_1\vec{v}_1 + c_2\vec{v}_2 + \cdots + c_n\vec{v}_n) \\ &= c_1A\vec{v}_1 + c_2A\vec{v}_2 + \cdots + c_nA\vec{v}_n \\ &= c_1PDP^{-1}\vec{v}_1 + c_2PDP^{-1}\vec{v}_2 + \cdots + c_nPDP^{-1}\vec{v}_n \\ &= PDP^{-1}(c_1\vec{v}_1 + c_2\vec{v}_2 + \cdots + c_n\vec{v}_n) \\ &= PDP^{-1}\vec{v}. \end{aligned}$$

Therefore, $A = PDP^{-1}$, and A is a diagonalizable matrix. □

Example 8.2.5: Diagonalize the matrix

$$A = \begin{bmatrix} 1 & 5 \\ 2 & -2 \end{bmatrix}$$

by finding the matrix P and D . Verify that $PDP^{-1} = A$.

Solution: In order to diagonalize the matrix, we look for the eigenvalues and eigenvectors. For the eigenvalues, we compute

$$\det(A - \lambda I) = (1 - \lambda)(-2 - \lambda) - 10 = \lambda^2 + \lambda - 12 = (\lambda + 4)(\lambda - 3).$$

So the eigenvalues of the matrix are -4 and 3 . For $\lambda = -4$, the system becomes

$$\begin{bmatrix} 5 & 5 \\ 2 & 2 \end{bmatrix} \vec{v} = 0$$

which gives an eigenvector $\vec{v} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

For $\lambda = 3$, we get the system

$$\begin{bmatrix} -2 & 5 \\ 2 & -5 \end{bmatrix} \vec{v} = 0$$

which gives an eigenvector $\vec{v} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$.

Therefore, we can write the matrices P and D from the proof of the previous theorem. P should contain the eigenvectors of A , and D should contain the eigenvalues in the same order as the eigenvectors in P . This means we can take

$$P = \begin{bmatrix} 1 & 5 \\ -1 & 2 \end{bmatrix} \quad D = \begin{bmatrix} -4 & 0 \\ 0 & 3 \end{bmatrix}.$$

To verify if this works, we need to compute P^{-1} . Using the formula for 2×2 matrices, we know that

$$P^{-1} = \frac{1}{2+5} \begin{bmatrix} 2 & -5 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{7} & \frac{-5}{7} \\ \frac{1}{7} & \frac{1}{7} \end{bmatrix}.$$

Then, we can compute

$$\begin{aligned} PDP^{-1} &= \begin{bmatrix} 1 & 5 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} -4 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} \frac{2}{7} & \frac{-5}{7} \\ \frac{1}{7} & \frac{1}{7} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 5 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} -\frac{8}{7} & \frac{20}{7} \\ \frac{3}{7} & \frac{3}{7} \end{bmatrix} \\ &= \begin{bmatrix} 1 & 5 \\ 2 & -2 \end{bmatrix} = A. \end{aligned}$$

So the diagonalization works as intended. □

The main question that needs to be answered is when is diagonalization possible? The main thing we need is this basis of eigenvectors from the matrix A . We know that this happens in two main situations:

- If the matrix A is symmetric (from the previous section), or,
- If the matrix A has all real and distinct eigenvalues, from § 3.6.

In both of those cases, we know we have the appropriate setup and can use this process to diagonalize the matrix. What happens in other situations? If there are complex eigenvalues (and complex eigenvectors) then it is possible to use them to diagonalize the matrix, but it brings complex numbers into the picture, which can make things much more complicated. The case where this really does not work is in the case of repeated eigenvalues, in particular, defective eigenvalues.

When we have defective eigenvalues, we can not get a full basis of eigenvectors. The best we can do here is to find as many eigenvectors as possible, and then fill out a basis with generalized eigenvectors. Let's see what this looks like in an example.

Example 8.2.6: Consider the matrix

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 3 \end{bmatrix}.$$

Find the eigenvalue, eigenvector, and generalized eigenvector. Form the matrix P with the eigenvector in the first column and generalized eigenvector in the second. Compute $P^{-1}AP$, which would be D in the previous cases.

Solution: The eigenvalues of this matrix can be found by

$$\det(A - \lambda I) = (1 - \lambda)(3 - \lambda) + 1 = \lambda^2 - 4\lambda + 4 = (\lambda - 2)^2.$$

For the only eigenvalue of $\lambda = 2$, the system becomes

$$\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \vec{v} = 0$$

so the eigenvector is $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. To find a generalized eigenvector, we want to find a solution to

$$\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix} \vec{w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

There are many vectors that can do this, one of which is $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Thus, the matrix P that we want to construct is

$$P = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix},$$

whose inverse is

$$P^{-1} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}.$$

Then, we can compute

$$\begin{aligned} P^{-1}AP &= \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 2 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}. \end{aligned}$$

So, the matrix we get is almost of the normal form, but has an extra 1 in the top right corner. The diagonal entries are also both 2, which is the same as the eigenvalue that these vectors came from. The form that we have here, with a 1 sitting between the repeated eigenvalue, is the best we can do in terms of getting close to a diagonalizable matrix when we have defective eigenvalues. This gives rise to the definition of a the Jordan form of a matrix.

Definition 8.2.4

A *Jordan Block* is an $m \times m$ square matrix with the same number in every diagonal entry and a 1 in each entry that is immediately above the diagonal.

A matrix J is in *Jordan form* if it consists of Jordan blocks along the diagonal of J and zeros everywhere else.

Example 8.2.7: An example of 1×1 , 2×2 and 3×3 Jordan blocks are

$$J_1 = [2] \quad J_2 = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix} \quad J_3 = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}.$$

The first of these has eigenvalue 2, the second has eigenvalue 3, and the third has eigenvalue -1 .

Remark 8.2.2: The number of 1s above the diagonal in a Jordan block indicates the defect of the corresponding eigenvalue. For the example above, we had a defect of 1, so we ended up with a block that looked like J_2 .

Example 8.2.8: The matrix J below is in Jordan form

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix}$$

$$J = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 3 & 1 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix}.$$

For this matrix, we have a 1-dimensional Jordan block of eigenvalue 1, which means there will be a single eigenvector with eigenvalue 1. Then there is a 2-dimensional Jordan block of eigenvalue 3, which means that there will be one eigenvector with eigenvalue 3 and a generalized eigenvector will be needed to make up the defect. Then there are two 1-dimensional Jordan blocks of eigenvalue 4, which means there will be two linearly independent eigenvalues with this eigenvalue. The characteristic polynomial of this matrix is

$$(\lambda - 1)(\lambda - 3)^2(\lambda - 4)^2.$$

Remark 8.2.3: Note that a diagonal matrix is in Jordan form, but all of the Jordan blocks are 1-dimensional. In that sense, these Jordan form matrices are an extension of diagonal matrices.

It is a theorem in linear algebra that, if we are allowed to take the eigenvalues to be complex numbers, then every matrix can be represented as PJP^{-1} , where P is an invertible matrix and J is a matrix in Jordan form. If we want to stick with real numbers, we have to be a bit more complicated with what our Jordan blocks look like, but it can be done.

8.2.5 Exercises

Exercise 8.2.2:* Determine if each of the following matrices are orthogonal.

a) $\begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}$

b) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

c) $\begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

d) $\begin{bmatrix} \frac{1}{9} & 0 & -\frac{4}{3\sqrt{2}} \\ \frac{2}{9} & \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} \\ \frac{2}{9} & -\frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} \end{bmatrix}$

Exercise 8.2.3:* Find all of the eigenvalues and eigenvectors of the matrix

$$\begin{bmatrix} 1 & 1 \\ -4 & 6 \end{bmatrix}.$$

Check if the eigenvectors for different eigenvalues are orthogonal. Does this make sense based on the original matrix?

Exercise 8.2.4: Find all of the eigenvalues and eigenvectors of the matrix

$$\begin{bmatrix} -3 & -6 \\ -6 & 13 \end{bmatrix}.$$

Check if the eigenvectors for different eigenvalues are orthogonal. Does this make sense based on the original matrix?

Exercise 8.2.5:* Find all of the eigenvalues and eigenvectors of the matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 2 \\ 0 & 2 & -1 \end{bmatrix}.$$

Check if the eigenvectors for different eigenvalues are orthogonal. Does this make sense based on the original matrix?

Exercise 8.2.6: Find all of the eigenvalues and eigenvectors of the matrix

$$\begin{bmatrix} -1 & -2 & 5 \\ -2 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix}.$$

Check if the eigenvectors for different eigenvalues are orthogonal. Does this make sense based on the original matrix?

Exercise 8.2.7: Diagonalize the matrix below using real numbers, if possible. To do so, find an invertible matrix P and diagonal matrix D so that $A = PDP^{-1}$. If this is not possible because of a repeated eigenvalue, state so, and find an invertible matrix P and Jordan Form matrix J so that $A = PJP^{-1}$.

$$A = \begin{bmatrix} -6 & -1 \\ 6 & -1 \end{bmatrix}.$$

Exercise 8.2.8:* Diagonalize the matrix below using real numbers, if possible. To do so, find an invertible matrix P and diagonal matrix D so that $A = PDP^{-1}$. If this is not possible because of a repeated eigenvalue, state so, and find an invertible matrix P and Jordan Form matrix J so that $A = PJP^{-1}$.

$$A = \begin{bmatrix} 7 & -8 \\ 3 & -3 \end{bmatrix}.$$

Exercise 8.2.9:* Diagonalize the matrix below using real numbers, if possible. To do so, find an invertible matrix P and diagonal matrix D so that $A = PDP^{-1}$. If this is not possible because of a repeated eigenvalue, state so, and find an invertible matrix P and Jordan Form matrix J so that $A = PJP^{-1}$.

$$A = \begin{bmatrix} 2 & -5 \\ 1 & 4 \end{bmatrix}.$$

Exercise 8.2.10:* Diagonalize the matrix below using real numbers, if possible. To do so, find an invertible matrix P and diagonal matrix D so that $A = PDP^{-1}$. If this is not possible because of a repeated eigenvalue, state so, and find an invertible matrix P and Jordan Form matrix J so that $A = PJP^{-1}$.

$$A = \begin{bmatrix} -2 & -2 \\ 8 & 6 \end{bmatrix}.$$

Exercise 8.2.11: Diagonalize the matrix below using real numbers, if possible. To do so, find an invertible matrix P and diagonal matrix D so that $A = PDP^{-1}$. If this is not possible because of a repeated eigenvalue, state so, and find an invertible matrix P and Jordan Form matrix J so that $A = PJP^{-1}$.

$$A = \begin{bmatrix} 2 & 0 & 4 \\ 3 & -1 & 4 \\ 0 & 0 & -4 \end{bmatrix}.$$

Exercise 8.2.12: Diagonalize the matrix below using real numbers, if possible. To do so, find an invertible matrix P and diagonal matrix D so that $A = PDP^{-1}$. If this is not possible

because of a repeated eigenvalue, state so, and find an invertible matrix P and Jordan Form matrix J so that $A = PJP^{-1}$.

$$A = \begin{bmatrix} -2 & 4 & 8 \\ -3 & -2 & 12 \\ 1 & 0 & -6 \end{bmatrix}.$$

Exercise 8.2.13: Diagonalize the matrix below using real numbers, if possible. To do so, find an invertible matrix P and diagonal matrix D so that $A = PDP^{-1}$. If this is not possible because of a repeated eigenvalue, state so, and find an invertible matrix P and Jordan Form matrix J so that $A = PJP^{-1}$.

$$A = \begin{bmatrix} -2 & 0 & 0 \\ -1 & 0 & -3 \\ -3 & 6 & -6 \end{bmatrix}.$$

8.3 Vector Spaces of Functions

Learning Objectives

After this section, you will be able to:

- Identify spaces of functions that are vector spaces,
- Compute inner products of functions using integrals, and
- Determine if functions are orthogonal with respect to this inner product.

The next step in our discussion of inner products and orthogonality is different places where this can apply. We have already talked about n component vectors and matrices, and the next example is function spaces.

8.3.1 Function Spaces

First, let's review the main properties of vector spaces that we defined in [Chapter 3](#). A (*real*) *vector space* is a set of objects V , called vectors, with two main properties

1. For any v and w in V , $v + w$ is also in V , and
2. For any v in V and α in \mathbb{R} , αv is in V .

The basic point here is that for something to be a vector space, we need to be able to add them together as well as multiply by real numbers, and these operations need to respect each other via the distributive law. We can do both of these things with functions that are defined on the same domain. For example, if we have two functions f and g defined on all real numbers and a real number α , then we can define two new functions $(f + g)$ and (αf) by

$$(f + g)(x) = f(x) + g(x) \quad (\alpha f)(x) = \alpha f(x).$$

These operations look a lot like the addition and scalar multiplication that we define on vectors, since those operations work component-wise and the function operations work “point-wise” at each value of x .

While this can be done for all functions, it isn't too useful unless we restrict what kinds of functions we actually care about.

Definition 8.3.1

Let $a < b$ be two real numbers. The set $C([a, b])$ is the set of all real-valued functions that are continuous on the interval $[a, b]$. The set $D([a, b])$ is the set of all real-valued functions that are differentiable on the interval $[a, b]$.

Both of these sets defined above are vector spaces! This result comes from work in Calculus 1 that shows that the sum of two continuous functions is continuous, and the same holds for differentiable functions. The proof of the fact that these are vector spaces is identical to what was done in Calculus 1 to prove these facts. These functions spaces are very large, as seen by the following example.

Example 8.3.1: Consider the space $C([-1, 1])$. This space contains all functions that are continuous on the interval $[-1, 1]$. This space contains $f_1(x) = x$, $f_2(x) = e^x$, $f_3(x) = \sin(x)$, $f_4(x) = \tan(x)$, $f_5(x) = \frac{1}{x-2}$, and $f_6(x) = |x|$. These last two functions do have asymptotes, but they are not within the interval $[1, 1]$. The function $g(x) = \frac{1}{x}$ is not in this space because it is discontinuous at $x = 0$. All of the f functions except for $f_6(x)$ are also in the space $D([-1, 1])$ because they are all differentiable on $[-1, 1]$ except for $|x|$ which is not differentiable at 0.

A lot of the ideas of vector spaces apply to these function spaces as well. The main one that is different is the idea of basis and dimension. Recall that a basis for a vector space V is a set of elements of V , denoted $B = \{v_1, v_2, \dots, v_n\}$ so that every element of V can be written as a linear combination of the elements of B . Furthermore, every basis of a given vector space has the same number of elements, which is the dimension of the vector space.

These functions spaces are distinctly different from the vector spaces discussed previously because these are infinite dimensional.

Example 8.3.2: Consider the space $C([-1, 1])$. Then all of the functions $f_0(x) = 1$, $f_1(x) = x$, ..., $f_k(x) = x^k$ are all in $C([-1, 1])$ and are all linearly independent.

Solution: All of the functions $f_k(x)$ for every k are all polynomials, so they are continuous on the whole real line, which also means they are continuous on $[-1, 1]$ and so are in $C([-1, 1])$. To establish linear independence, we can't use matrices, but instead need to go back to the definition of linear independence.

First, we check if $f_0(x) = 1$ and $f_1(x) = x$ are linearly independent. For this, we need to determine if it is possible to pick c_0 and c_1 so that

$$c_0 f_0 + c_1 f_1 = 0,$$

or in this particular case

$$c_0 + c_1 x = 0.$$

Now, for this to equal zero in the sense of functions in $C([-1, 1])$, we need this expression to be zero for **all** x values in the interval $[-1, 1]$, not just some of them. We know that the graph of $c_1 x + c_0$ is a straight line, so the only way it can be zero for all x in $[-1, 1]$ is if it is the line $y = 0$, so both c_1 and c_0 are zero. Therefore, we know that f_0 and f_1 are linearly independent.

Now, we want to include $f_2(x) = x^2$ in this set. To determine this linear independence, we need to determine if there are coefficients c_0 , c_1 and c_2 so that

$$c_0 + c_1 x + c_2 x^2 = 0$$

for **all** x in $[-1, 1]$. Since the graph of $c_2 x^2 + c_1 x + c_0$ is a parabola, it can have at most two zeros. The only way this can be zero for all values of x is if it is identically zero, so we need c_0 , c_1 , and c_2 equal to zero, so these three functions are linearly independent.

The same process continues for all k . If we pick some upper level k , then we need to find constants so that

$$c_0 + c_1 x + c_2 x^2 + \cdots + c_k x^k = 0$$

for all x in $[-1, 1]$. However, a polynomial of degree k can have at most k zeros, and since there are more than k points in $[-1, 1]$ (there are infinitely many), for this to be zero everywhere, we need to have all of the constants zero, and so the functions are linearly independent. \square

This tells us that the space $C([-1, 1])$ has an infinite set, $\{1, x, x^2, \dots\}$ of linearly independent functions. This means that there is no finite set of functions that span the entire space, which means that the space is infinite dimensional. This is very different from the finite dimensional vector spaces that we discussed previously. A lot of the ideas will carry through, but we will always have to keep in mind that we don't have bases or matrices to fall back on.

8.3.2 Inner Products on Function Spaces

Since function spaces like $C([a, b])$ and $D([a, b])$ are vector spaces, we can look into trying to define an inner product on them. In terms of vector spaces, we formed an inner product by taking two vectors, multiplying each of their individual components together and adding them up. Components for vectors are related to function values at particular points, so we can try to do this same process on functions, and it results in an integral.

Definition 8.3.2

Let f and g be two functions in either $C([a, b])$ or $D([a, b])$. An inner product on this space can be defined by

$$\langle f, g \rangle = \int_a^b f(x)g(x) dx.$$

As discussed in § 8.1, any inner product needs to satisfy the following properties:

- (i) $\langle f, f \rangle \geq 0$, and $\langle f, f \rangle = 0$ if and only if $f = 0$,
- (ii) $\langle f, g \rangle = \langle g, f \rangle$,
- (iii) $\langle af, g \rangle = \langle f, ag \rangle = a\langle f, g \rangle$,
- (iv) $\langle f + g, h \rangle = \langle f, h \rangle + \langle g, h \rangle$ and $\langle f, g + h \rangle = \langle f, g \rangle + \langle f, h \rangle$.

All of these properties follow from linearity of the integral and commutativity of standard products. The trickiest one of them is the fact that $\langle f, f \rangle \geq 0$, which comes from the fact that f^2 is always a positive function for every f , and so the integral is positive.

Exercise 8.3.1: Write out all of these properties in terms of integrals and verify them.

Example 8.3.3: Consider the space $C([-1, 1])$. Compute the inner product $\langle x, x^2 \rangle$ and $\langle x, e^x \rangle$.

Solution: Since the space is defined on $[-1, 1]$, the integrals that we need to compute should be done over this range. Then, we have that

$$\langle x, x^2 \rangle = \int_{-1}^1 x(x^2) dx = \int_{-1}^1 x^3 dx = \frac{x^4}{4} \Big|_{-1}^1 = 0$$

and

$$\begin{aligned}
 \langle x, e^x \rangle &= \int_{-1}^1 xe^x \, dx \\
 &= xe^x - \int_{-1}^1 e^x \, dx \\
 &= xe^x - e^x \Big|_{-1}^1 \\
 &= (e - e) - (-e^{-1} - e^{-1}) = 2e^{-1}.
 \end{aligned}$$

□

Remark 8.3.1: The interval on which these spaces are defined is critical and can change the value of the inner product. For example, if we had defined these on the space $C([0, 2])$, then the first pair of functions would have inner product equal to 4, and the second would have inner product equal to $3e^4 - 1$.

Orthogonal Functions

Now that we have an inner product, we can talk about orthogonality of functions. In some sense, we can talk about angles between functions, but since there isn't a direct geometric interpretation, the only thing that really matters is functions that are perpendicular.

Definition 8.3.3

Let f and g be two continuous functions on $[a, b]$. We say that these functions are *orthogonal* if

$$\langle f, g \rangle = \int_a^b f(x)g(x) \, dx = 0.$$

Above we computed that, considering the space $C([-1, 1])$, $\langle x, x^2 \rangle = 0$. Therefore, in this space, the functions x and x^2 are orthogonal, but x and e^x are not. However, the remark shows that this is dependent on the interval where these spaces are defined, because on $[0, 2]$, x and x^2 are not orthogonal.

Example 8.3.4: Consider $C([0, 2])$. Are the functions $f_1(x) = x^2$ and $f_2(x) = x^3$ orthogonal? Determine a value of c so that f_1 and $g(x) = x^3 - cx^2$ are orthogonal.

Solution: For the first part, we compute that

$$\langle x^2, x^3 \rangle = \int_0^2 (x^2)(x^3) \, dx = \int_0^2 x^5 \, dx = \frac{x^6}{6} \Big|_0^2 = \frac{64}{6} = \frac{32}{3}.$$

Therefore, these functions are not orthogonal. For the second part, we compute the inner

product

$$\begin{aligned}\langle x^2, x^3 - cx^2 \rangle &= \int_0^2 x^2(x^3 - cx^2) \, dx \\ &= \int_0^2 x^5 - cx^4 \, dx \\ &= \frac{x^6}{6} - c\frac{x^5}{5} \Big|_0^2 \\ &= \frac{32}{3} - c\frac{32}{5} = 32\left(\frac{1}{3} - \frac{c}{5}\right).\end{aligned}$$

If we want these functions to be orthogonal, we want this to evaluate to 0. Thus, we need $\frac{1}{3} = \frac{c}{5}$, so that $c = \frac{5}{3}$. □

A common way to discuss these orthogonal functions is by talking about classes of functions. Finding one or two orthogonal functions isn't entirely useful, but if we can find a large collection of them, that has uses in the future.

Example 8.3.5: Are the monomials $S = \{1, x, x^2, \dots\}$ orthogonal on $[-1, 1]$?

Solution: We can compute these inner products directly. Any function pulled from this set will be of the form x^m for some positive integer m . Thus, the inner product will be

$$\begin{aligned}\langle x^m, x^n \rangle &= \int_{-1}^1 (x^m)(x^n) \, dx \\ &= \int_{-1}^1 x^{m+n} \, dx \\ &= \frac{x^{m+n+1}}{m+n+1} \Big|_{-1}^1 = \frac{1}{m+n+1} - \frac{(-1)^{m+n+1}}{m+n+1}.\end{aligned}$$

So, if $m + n + 1$ is even, then this difference will be zero, and so the functions will be orthogonal. If $m + n + 1$ is odd, then the difference will not be zero, and they will not be orthogonal. Thus, for any functions in this set, x^m and x^n are orthogonal if exactly one of m and n is even. This means that the **entire** set S is not orthogonal, because there are pairs of functions in that set that are not orthogonal. □

The most famous set of orthogonal functions is trigonometric functions, which will be walked through in some of the exercises. These functions will form the basis for Fourier Series, which will be discussed in [Chapter 9](#).

8.3.3 Exercises

Exercise 8.3.2: Consider the space $C([0, 1])$. Compute the following inner products.

- a) $\langle x, x^3 \rangle$
- b) $\langle e^x, x + 1 \rangle$
- c) $\langle e^x, \sin(\pi x) \rangle$
- d) $\langle x^2 + 1, x - 3 \rangle$.

Exercise 8.3.3: Are the functions $\cos(x)$ and $\sin(x)$ orthogonal in $C([-\pi, \pi])$? What about in $C([0, \pi])$?

Exercise 8.3.4: Are the functions $\cos(x)$ and $\cos(2x)$ orthogonal in $C([-\pi, \pi])$? What about in $C([0, \pi])$? **Hint:** The product to sum formulas for trigonometric functions will be useful here, in particular,

$$\cos(A)\cos(B) = \frac{1}{2}(\cos(A+B) + \cos(A-B)).$$

Exercise 8.3.5: Consider the set of functions $S = \{\sin(x), \sin(2x), \sin(3x), \dots\}$ and $C = \{\cos(x), \cos(2x), \cos(3x), \dots\}$ as functions in $C([-\pi, \pi])$.

- a) Is S an orthogonal set in this space? You will want to start by picking two functions in this set, which will be $\sin(nx)$ and $\sin(mx)$, and computing the inner product.
- b) Is C an orthogonal set in this space?
- c) If we combine C and S , do we still get an orthogonal set? After establishing the first two parts, we only need to look at the inner product with one function from C and one function from S .
- d) Are these functions in C and S orthogonal to the constant function 1?

For this, the product-to-sum identities will be useful.

$$\begin{aligned}\cos(A)\cos(B) &= \frac{1}{2}(\cos(A+B) + \cos(A-B)) \\ \sin(A)\cos(B) &= \frac{1}{2}(\sin(A+B) + \sin(A-B)) \\ \sin(A)\sin(B) &= \frac{1}{2}(\cos(A-B) - \cos(A+B))\end{aligned}$$

Chapter 9

Fourier series

9.1 Boundary value problems

Attribution: [JL], §4.1.

Learning Objectives

After this section, you will be able to:

- Analyze two-point boundary value problems on a given interval,
- Find the eigenvalues and eigenfunctions for a boundary value problem, and
- Apply these techniques to an oscillating string problem.

9.1.1 Boundary value problems

Before we tackle the Fourier series, we study the so-called *boundary value problems* (or *endpoint problems*). Consider

$$x'' + \lambda x = 0, \quad x(a) = 0, \quad x(b) = 0,$$

for some constant λ , where $x(t)$ is defined for t in the interval $[a, b]$. Previously we specified the value of the solution and its derivative at a single point. Now we specify the value of the solution at two different points. As $x = 0$ is a solution, existence of solutions is not a problem. Uniqueness of solutions is another issue. The general solution to $x'' + \lambda x = 0$ has two arbitrary constants[†]. It is, therefore, natural (but wrong) to believe that requiring two conditions guarantees a unique solution.

Example 9.1.1: Take $\lambda = 1$, $a = 0$, $b = \pi$. That is,

$$x'' + x = 0, \quad x(0) = 0, \quad x(\pi) = 0.$$

Then $x = \sin t$ is another solution (besides $x = 0$) satisfying both boundary conditions. There are more. Write down the general solution of the differential equation, which is

[†]See subsection 0.1.4 on page 16 or Example 2.2.1 on page 114.

$x = A \cos t + B \sin t$. The condition $x(0) = 0$ forces $A = 0$. Letting $x(\pi) = 0$ does not give us any more information as $x = B \sin t$ already satisfies both boundary conditions. Hence, there are infinitely many solutions of the form $x = B \sin t$, where B is an arbitrary constant.

Example 9.1.2: On the other hand, consider $\lambda = 2$. That is,

$$x'' + 2x = 0, \quad x(0) = 0, \quad x(\pi) = 0.$$

Then the general solution is $x = A \cos(\sqrt{2}t) + B \sin(\sqrt{2}t)$. Letting $x(0) = 0$ still forces $A = 0$. We apply the second condition to find $0 = x(\pi) = B \sin(\sqrt{2}\pi)$. As $\sin(\sqrt{2}\pi) \neq 0$ we obtain $B = 0$. Therefore $x = 0$ is the unique solution to this problem.

What is going on? We will be interested in finding which constants λ allow a nonzero solution, and we will be interested in finding those solutions. This problem is an analogue of finding eigenvalues and eigenvectors of matrices.

9.1.2 Eigenvalue problems

For basic Fourier series theory we will need the following three eigenvalue problems.

$$x'' + \lambda x = 0, \quad x(a) = 0, \quad x(b) = 0, \quad (9.1)$$

$$x'' + \lambda x = 0, \quad x'(a) = 0, \quad x'(b) = 0, \quad (9.2)$$

and

$$x'' + \lambda x = 0, \quad x(a) = x(b), \quad x'(a) = x'(b). \quad (9.3)$$

Definition 9.1.1

A number λ is called an *eigenvalue* of (9.1) (resp. (9.2) or (9.3)) if and only if there exists a nonzero (not identically zero) solution to (9.1) (resp. (9.2) or (9.3)) given that specific λ . A nonzero solution is called a corresponding *eigenfunction*.

Note the similarity to eigenvalues and eigenvectors of matrices. The similarity is not just coincidental. If we think of the equations as differential operators, then we are doing the same exact thing. Think of a function $x(t)$ as a vector with infinitely many components (one for each t). Let $L = -\frac{d^2}{dt^2}$ be the linear operator. Then the eigenvalue/eigenfunction pair should be λ and nonzero x such that $Lx = \lambda x$. In other words, we are looking for nonzero functions x satisfying certain endpoint conditions that solve $(L - \lambda)x = 0$. A lot of the formalism from linear algebra still applies here, though we will not pursue this line of reasoning too far.

Example 9.1.3: Let us find the eigenvalues and eigenfunctions of

$$x'' + \lambda x = 0, \quad x(0) = 0, \quad x(\pi) = 0.$$

Solution: We have to handle the cases $\lambda > 0$, $\lambda = 0$, $\lambda < 0$ separately. First suppose that $\lambda > 0$. Then the general solution to $x'' + \lambda x = 0$ is

$$x = A \cos(\sqrt{\lambda}t) + B \sin(\sqrt{\lambda}t).$$

The condition $x(0) = 0$ implies immediately $A = 0$. Next

$$0 = x(\pi) = B \sin(\sqrt{\lambda} \pi).$$

If B is zero, then x is not a nonzero solution. So to get a nonzero solution we must have that $\sin(\sqrt{\lambda} \pi) = 0$. Hence, $\sqrt{\lambda} \pi$ must be an integer multiple of π . In other words, $\sqrt{\lambda} = k$ for a positive integer k . Hence the positive eigenvalues are k^2 for all integers $k \geq 1$. Corresponding eigenfunctions can be taken as $x = \sin(kt)$. Just like for eigenvectors, constant multiples of an eigenfunction are also eigenfunctions, so we only need to pick one.

Now suppose that $\lambda = 0$. In this case the equation is $x'' = 0$, and its general solution is $x = At + B$. The condition $x(0) = 0$ implies that $B = 0$, and $x(\pi) = 0$ implies that $A = 0$. This means that $\lambda = 0$ is *not* an eigenvalue.

Finally, suppose that $\lambda < 0$. In this case we have the general solution*

$$x = A \cosh(\sqrt{-\lambda} t) + B \sinh(\sqrt{-\lambda} t).$$

Letting $x(0) = 0$ implies that $A = 0$ (recall $\cosh 0 = 1$ and $\sinh 0 = 0$). So our solution must be $x = B \sinh(\sqrt{-\lambda} t)$ and satisfy $x(\pi) = 0$. This is only possible if B is zero. Why? Because $\sinh \xi$ is only zero when $\xi = 0$. You should plot \sinh to see this fact. We can also see this from the definition of \sinh . We get $0 = \sinh \xi = \frac{e^\xi - e^{-\xi}}{2}$. Hence $e^\xi = e^{-\xi}$, which implies $\xi = -\xi$ and that is only true if $\xi = 0$. So there are no negative eigenvalues.

In summary, the eigenvalues and corresponding eigenfunctions are

$$\lambda_k = k^2 \quad \text{with an eigenfunction} \quad x_k = \sin(kt) \quad \text{for all integers } k \geq 1.$$

□

Example 9.1.4: Let us compute the eigenvalues and eigenfunctions of

$$x'' + \lambda x = 0, \quad x'(0) = 0, \quad x'(\pi) = 0.$$

Solution: Again we have to handle the cases $\lambda > 0$, $\lambda = 0$, $\lambda < 0$ separately. First suppose that $\lambda > 0$. The general solution to $x'' + \lambda x = 0$ is $x = A \cos(\sqrt{\lambda} t) + B \sin(\sqrt{\lambda} t)$. So

$$x' = -A\sqrt{\lambda} \sin(\sqrt{\lambda} t) + B\sqrt{\lambda} \cos(\sqrt{\lambda} t).$$

The condition $x'(0) = 0$ implies immediately $B = 0$. Next

$$0 = x'(\pi) = -A\sqrt{\lambda} \sin(\sqrt{\lambda} \pi).$$

Again A cannot be zero if λ is to be an eigenvalue, and $\sin(\sqrt{\lambda} \pi)$ is only zero if $\sqrt{\lambda} = k$ for a positive integer k . Hence the positive eigenvalues are again k^2 for all integers $k \geq 1$. And the corresponding eigenfunctions can be taken as $x = \cos(kt)$.

Now suppose that $\lambda = 0$. In this case the equation is $x'' = 0$ and the general solution is $x = At + B$ so $x' = A$. The condition $x'(0) = 0$ implies that $A = 0$. The condition $x'(\pi) = 0$

*Recall that $\cosh s = \frac{1}{2}(e^s + e^{-s})$ and $\sinh s = \frac{1}{2}(e^s - e^{-s})$. As an exercise try the computation with the general solution written as $x = Ae^{\sqrt{-\lambda} t} + Be^{-\sqrt{-\lambda} t}$ (for different A and B of course).

also implies $A = 0$. Hence B could be anything (let us take it to be 1). So $\lambda = 0$ is an eigenvalue and $x = 1$ is a corresponding eigenfunction.

Finally, let $\lambda < 0$. In this case the general solution is $x = A \cosh(\sqrt{-\lambda} t) + B \sinh(\sqrt{-\lambda} t)$ and

$$x' = A\sqrt{-\lambda} \sinh(\sqrt{-\lambda} t) + B\sqrt{-\lambda} \cosh(\sqrt{-\lambda} t).$$

We have already seen (with roles of A and B switched) that for this expression to be zero at $t = 0$ and $t = \pi$, we must have $A = B = 0$. Hence there are no negative eigenvalues.

In summary, the eigenvalues and corresponding eigenfunctions are

$$\lambda_k = k^2 \quad \text{with an eigenfunction} \quad x_k = \cos(kt) \quad \text{for all integers } k \geq 1,$$

and there is another eigenvalue

$$\lambda_0 = 0 \quad \text{with an eigenfunction} \quad x_0 = 1. \quad \square$$

The following problem is the one that leads to the general Fourier series.

Example 9.1.5: Let us compute the eigenvalues and eigenfunctions of

$$x'' + \lambda x = 0, \quad x(-\pi) = x(\pi), \quad x'(-\pi) = x'(\pi).$$

We have not specified the values or the derivatives at the endpoints, but rather that they are the same at the beginning and at the end of the interval.

Solution: Let us skip $\lambda < 0$. The computations are the same as before, and again we find that there are no negative eigenvalues.

For $\lambda = 0$, the general solution is $x = At + B$. The condition $x(-\pi) = x(\pi)$ implies that $A = 0$ ($A\pi + B = -A\pi + B$ implies $A = 0$). The second condition $x'(-\pi) = x'(\pi)$ says nothing about B and hence $\lambda = 0$ is an eigenvalue with a corresponding eigenfunction $x = 1$.

For $\lambda > 0$ we get that $x = A \cos(\sqrt{\lambda} t) + B \sin(\sqrt{\lambda} t)$. Now

$$\underbrace{A \cos(-\sqrt{\lambda} \pi) + B \sin(-\sqrt{\lambda} \pi)}_{x(-\pi)} = \underbrace{A \cos(\sqrt{\lambda} \pi) + B \sin(\sqrt{\lambda} \pi)}_{x(\pi)}.$$

We remember that $\cos(-\theta) = \cos(\theta)$ and $\sin(-\theta) = -\sin(\theta)$. Therefore,

$$A \cos(\sqrt{\lambda} \pi) - B \sin(\sqrt{\lambda} \pi) = A \cos(\sqrt{\lambda} \pi) + B \sin(\sqrt{\lambda} \pi).$$

Hence either $B = 0$ or $\sin(\sqrt{\lambda} \pi) = 0$. Similarly (exercise) if we differentiate x and plug in the second condition we find that $A = 0$ or $\sin(\sqrt{\lambda} \pi) = 0$. Therefore, unless we want A and B to both be zero (which we do not) we must have $\sin(\sqrt{\lambda} \pi) = 0$. Hence, $\sqrt{\lambda}$ is an integer and the eigenvalues are yet again $\lambda = k^2$ for an integer $k \geq 1$. In this case, however, $x = A \cos(kt) + B \sin(kt)$ is an eigenfunction for any A and any B . So we have two linearly independent eigenfunctions $\sin(kt)$ and $\cos(kt)$. Remember that for a matrix we can also have two eigenvectors corresponding to a single eigenvalue if the eigenvalue is repeated.

In summary, the eigenvalues and corresponding eigenfunctions are

$$\lambda_k = k^2 \quad \text{with eigenfunctions} \quad \cos(kt) \quad \text{and} \quad \sin(kt) \quad \text{for all integers } k \geq 1,$$

$$\lambda_0 = 0 \quad \text{with an eigenfunction} \quad x_0 = 1. \quad \square$$

9.1.3 Orthogonality of eigenfunctions

Something that will be very useful in the next section is the *orthogonality* property of the eigenfunctions. This is an analogue of the following fact about eigenvectors of a matrix. A matrix is called *symmetric* if $A = A^T$ (it is equal to its transpose). *Eigenvectors for two distinct eigenvalues of a symmetric matrix are orthogonal*, which was shown in § 8.2. The differential operators we are dealing with act much like a symmetric matrix. We, therefore, get the following theorem.

Theorem 9.1.1

Suppose that $x_1(t)$ and $x_2(t)$ are two eigenfunctions of the problem (9.1), (9.2) or (9.3) for two different eigenvalues λ_1 and λ_2 . Then they are *orthogonal* in the sense that

$$\int_a^b x_1(t)x_2(t) dt = 0.$$

The terminology comes from the fact that the integral is a type of inner product. We will expand on this in the next section. The theorem has a very short, elegant, and illuminating proof so let us give it here. First, we have the following two equations.

$$x_1'' + \lambda_1 x_1 = 0 \quad \text{and} \quad x_2'' + \lambda_2 x_2 = 0.$$

Multiply the first by x_2 and the second by x_1 and subtract to get

$$(\lambda_1 - \lambda_2)x_1x_2 = x_2''x_1 - x_2x_1''.$$

Now integrate both sides of the equation:

$$\begin{aligned} (\lambda_1 - \lambda_2) \int_a^b x_1x_2 dt &= \int_a^b x_2''x_1 - x_2x_1'' dt \\ &= \int_a^b \frac{d}{dt} (x_2'x_1 - x_2x_1') dt \\ &= \left[x_2'x_1 - x_2x_1' \right]_{t=a}^b = 0. \end{aligned}$$

The last equality holds because of the boundary conditions. For example, if we consider (9.1) we have $x_1(a) = x_1(b) = x_2(a) = x_2(b) = 0$ and so $x_2'x_1 - x_2x_1'$ is zero at both a and b . As $\lambda_1 \neq \lambda_2$, the theorem follows.

Exercise 9.1.1 (easy): *Finish the proof of the theorem (check the last equality in the proof) for the cases (9.2) and (9.3).*

The function $\sin(nt)$ is an eigenfunction for the problem $x'' + \lambda x = 0$, $x(0) = 0$, $x(\pi) = 0$. Hence for positive integers n and m we have the integrals

$$\int_0^\pi \sin(mt) \sin(nt) dt = 0, \quad \text{when } m \neq n.$$

Similarly,

$$\int_0^\pi \cos(mt) \cos(nt) dt = 0, \quad \text{when } m \neq n, \quad \text{and} \quad \int_0^\pi \cos(nt) dt = 0.$$

And finally we also get

$$\begin{aligned} \int_{-\pi}^\pi \sin(mt) \sin(nt) dt &= 0, \quad \text{when } m \neq n, \quad \text{and} \quad \int_{-\pi}^\pi \sin(nt) dt = 0, \\ \int_{-\pi}^\pi \cos(mt) \cos(nt) dt &= 0, \quad \text{when } m \neq n, \quad \text{and} \quad \int_{-\pi}^\pi \cos(nt) dt = 0, \end{aligned}$$

and

$$\int_{-\pi}^\pi \cos(mt) \sin(nt) dt = 0 \quad (\text{even if } m = n).$$

9.1.4 Fredholm alternative

We now touch on a very useful theorem in the theory of differential equations. The theorem holds in a more general setting than we are going to state it, but for our purposes the following statement is sufficient.

Theorem 9.1.2 (Fredholm alternative)

Exactly one of the following statements holds. Either

$$x'' + \lambda x = 0, \quad x(a) = 0, \quad x(b) = 0 \tag{9.4}$$

has a nonzero solution, or

$$x'' + \lambda x = f(t), \quad x(a) = 0, \quad x(b) = 0 \tag{9.5}$$

has a unique solution for every function f continuous on $[a, b]$.

The theorem* is also true for the other types of boundary conditions we considered. The theorem means that if λ is not an eigenvalue, the nonhomogeneous equation (9.5) has a unique solution for every right-hand side. On the other hand if λ is an eigenvalue, then (9.5) need not have a solution for every f , and furthermore, even if it happens to have a solution, the solution is not unique.

We also want to reinforce the idea here that linear differential operators have much in common with matrices. So it is no surprise that there is a finite-dimensional version of Fredholm alternative for matrices as well. Let A be an $n \times n$ matrix. The Fredholm alternative then states that either $(A - \lambda I)\vec{x} = \vec{0}$ has a nontrivial solution, or $(A - \lambda I)\vec{x} = \vec{b}$ has a unique solution for every \vec{b} .

A lot of intuition from linear algebra can be applied to linear differential operators, but one must be careful of course. For example, one difference we have already seen is that in general a differential operator will have infinitely many eigenvalues, while a matrix has only finitely many.

*Named after the Swedish mathematician Erik Ivar Fredholm (1866–1927).

9.1.5 Application

Let us consider a physical application of an endpoint problem. Suppose we have a tightly stretched quickly spinning elastic string or rope of uniform linear density ρ , for example in kg/m . Let us put this problem into the xy -plane and both x and y are in meters. The x -axis represents the position on the string. The string rotates at angular velocity ω , in radians/s . Imagine that the whole xy -plane rotates at angular velocity ω . This way, the string stays in this xy -plane and y measures its deflection from the equilibrium position, $y = 0$, on the x -axis. Hence the graph of y gives the shape of the string. We consider an ideal string with no volume, just a mathematical curve. We suppose the tension on the string is a constant T in Newtons. Assuming that the deflection is small, we can use Newton's second law (let us skip the derivation) to get the equation

$$Ty'' + \rho\omega^2y = 0.$$

To check the units notice that the units of y'' are m/m^2 , as the derivative is in terms of x .

Let L be the length of the string (in meters) and the string is fixed at the beginning and end points. Hence, $y(0) = 0$ and $y(L) = 0$. See [Figure 9.1](#).

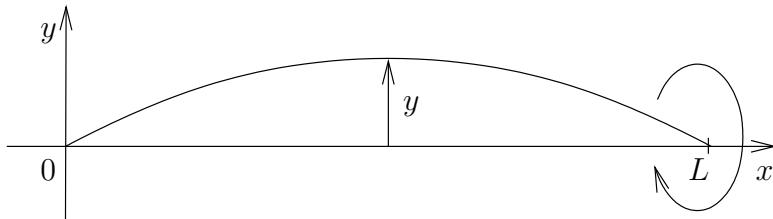


Figure 9.1: Whirling string.

We rewrite the equation as $y'' + \frac{\rho\omega^2}{T}y = 0$. The setup is similar to [Example 9.1.3](#) on page 448, except for the interval length being L instead of π . We are looking for eigenvalues of $y'' + \lambda y = 0$, $y(0) = 0$, $y(L) = 0$ where $\lambda = \frac{\rho\omega^2}{T}$. As before there are no nonpositive eigenvalues. With $\lambda > 0$, the general solution to the equation is $y = A \cos(\sqrt{\lambda}x) + B \sin(\sqrt{\lambda}x)$. The condition $y(0) = 0$ implies that $A = 0$ as before. The condition $y(L) = 0$ implies that $\sin(\sqrt{\lambda}L) = 0$ and hence $\sqrt{\lambda}L = k\pi$ for some integer $k > 0$, so

$$\frac{\rho\omega^2}{T} = \lambda = \frac{k^2\pi^2}{L^2}.$$

What does this say about the shape of the string? It says that for all parameters ρ , ω , T not satisfying the equation above, the string is in the equilibrium position, $y = 0$. When $\frac{\rho\omega^2}{T} = \frac{k^2\pi^2}{L^2}$, then the string will “pop out” some distance B . We cannot compute B with the information we have.

Let us assume that ρ and T are fixed and we are changing ω . For most values of ω the string is in the equilibrium state. When the angular velocity ω hits a value $\omega = \frac{k\pi\sqrt{T}}{L\sqrt{\rho}}$, then the string pops out and has the shape of a sin wave crossing the x -axis $k - 1$ times between

the end points. For example, at $k = 1$, the string does not cross the x -axis and the shape looks like in [Figure 9.1](#) on the preceding page. On the other hand, when $k = 3$ the string crosses the x -axis 2 times, see [Figure 9.2](#). When ω changes again, the string returns to the equilibrium position. The higher the angular velocity, the more times it crosses the x -axis when it is popped out.

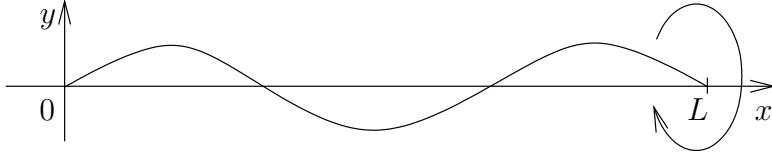


Figure 9.2: Whirling string at the third eigenvalue ($k = 3$).

For another example, if you have a spinning jump rope (then $k = 1$ as it is completely “popped out”) and you pull on the ends to increase the tension, then the velocity also increases for the rope to stay “popped out”.

9.1.6 Exercises

Exercise 9.1.2:* Consider a spinning string of length 2 and linear density 0.1 and tension 3. Find smallest angular velocity when the string pops out.

Hint for the following exercises: Note that when $\lambda > 0$, then $\cos(\sqrt{\lambda}(t - a))$ and $\sin(\sqrt{\lambda}(t - a))$ are also solutions of the homogeneous equation.

Exercise 9.1.3: Compute all eigenvalues and eigenfunctions of $x'' + \lambda x = 0$, $x(a) = 0$, $x(b) = 0$ (assume $a < b$).

Exercise 9.1.4: Compute all eigenvalues and eigenfunctions of $x'' + \lambda x = 0$, $x'(a) = 0$, $x'(b) = 0$ (assume $a < b$).

Exercise 9.1.5:* Suppose $x'' + \lambda x = 0$ and $x(0) = 1$, $x(1) = 1$. Find all λ for which there is more than one solution. Also find the corresponding solutions (only for the eigenvalues).

Exercise 9.1.6: Compute all eigenvalues and eigenfunctions of $x'' + \lambda x = 0$, $x'(a) = 0$, $x(b) = 0$ (assume $a < b$).

Exercise 9.1.7: Compute all eigenvalues and eigenfunctions of $x'' + \lambda x = 0$, $x(a) = x(b)$, $x'(a) = x'(b)$ (assume $a < b$).

Exercise 9.1.8: Suppose $x'' + x = 0$ and $x(0) = 0$, $x'(\pi) = 1$. Find all the solution(s) if any exist.

Exercise 9.1.9: We skipped the case of $\lambda < 0$ for the boundary value problem $x'' + \lambda x = 0$, $x(-\pi) = x(\pi)$, $x'(-\pi) = x'(\pi)$. Finish the calculation and show that there are no negative eigenvalues.

Exercise 9.1.10:* Consider $x' + \lambda x = 0$ and $x(0) = 0$, $x(1) = 0$. Why does it not have any eigenvalues? Why does any first order equation with two endpoint conditions such as above have no eigenvalues?

Exercise 9.1.11 (challenging):* Suppose $x''' + \lambda x = 0$ and $x(0) = 0$, $x'(0) = 0$, $x(1) = 0$. Suppose that $\lambda > 0$. Find an equation that all such eigenvalues must satisfy. Hint: Note that $-\sqrt[3]{\lambda}$ is a root of $r^3 + \lambda = 0$.

9.2 The trigonometric series

Attribution: [JL], §4.2.

Learning Objectives

After this section, you will be able to:

- Find and compute periodic extensions of functions and
- Find the Fourier series for a given function.

9.2.1 Periodic functions and motivation

As motivation for studying Fourier series, suppose we have the problem

$$x'' + \omega_0^2 x = f(t), \quad (9.6)$$

for some periodic function $f(t)$. We already solved

$$x'' + \omega_0^2 x = F_0 \cos(\omega t). \quad (9.7)$$

One way to solve (9.6) is to decompose $f(t)$ as a sum of cosines (and sines) and then solve many problems of the form (9.7). We then use the principle of superposition, to sum up all the solutions we got to get a solution to (9.6).

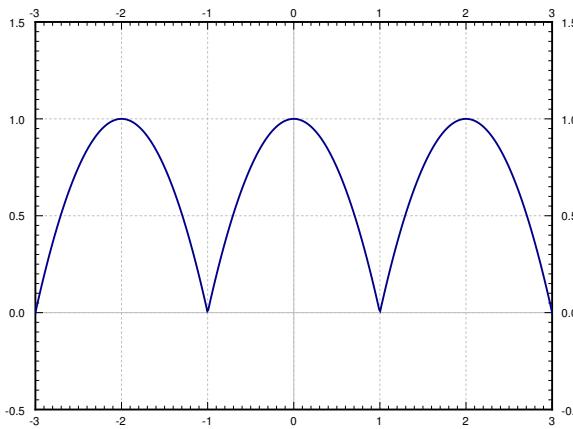
Before we proceed, let us talk a little bit more in detail about periodic functions. A function is said to be *periodic* with period P if $f(t) = f(t + P)$ for all t . For brevity we say $f(t)$ is P -periodic. Note that a P -periodic function is also $2P$ -periodic, $3P$ -periodic and so on. For example, $\cos(t)$ and $\sin(t)$ are 2π -periodic. So are $\cos(kt)$ and $\sin(kt)$ for all integers k . The constant functions are an extreme example. They are periodic for any period (exercise).

Normally we start with a function $f(t)$ defined on some interval $[-L, L]$, and we want to *extend* $f(t)$ *periodically* to make it a $2L$ -periodic function. We do this extension by defining a new function $F(t)$ such that for t in $[-L, L]$, $F(t) = f(t)$. For t in $[L, 3L]$, we define $F(t) = f(t - 2L)$, for t in $[-3L, -L]$, $F(t) = f(t + 2L)$, and so on. To make that work we needed $f(-L) = f(L)$. We could have also started with f defined only on the half-open interval $(-L, L]$ and then define $f(-L) = f(L)$.

Example 9.2.1: Define $f(t) = 1 - t^2$ on $[-1, 1]$. Now extend $f(t)$ periodically to a 2-periodic function. See [Figure 9.3](#) on the facing page.

You should be careful to distinguish between $f(t)$ and its extension. A common mistake is to assume that a formula for $f(t)$ holds for its extension. It can be confusing when the formula for $f(t)$ is periodic, but with perhaps a different period.

Exercise 9.2.1: Define $f(t) = \cos t$ on $[-\pi/2, \pi/2]$. Take the π -periodic extension and sketch its graph. How does it compare to the graph of $\cos t$?

Figure 9.3: Periodic extension of the function $1 - t^2$.

9.2.2 Inner product and eigenvector decomposition

Suppose we have a *symmetric matrix*, that is $A^T = A$. As we remarked before, eigenvectors of A are then orthogonal. Here the word *orthogonal* means that if \vec{v} and \vec{w} are two eigenvectors of A for distinct eigenvalues, then $\langle \vec{v}, \vec{w} \rangle = 0$. In this case the inner product $\langle \vec{v}, \vec{w} \rangle$ is the *dot product*, which can be computed as $\vec{v}^T \vec{w}$.

To decompose a vector \vec{v} in terms of mutually orthogonal vectors \vec{w}_1 and \vec{w}_2 we write

$$\vec{v} = a_1 \vec{w}_1 + a_2 \vec{w}_2.$$

Let us find the formula for a_1 and a_2 . First let us compute

$$\langle \vec{v}, \vec{w}_1 \rangle = \langle a_1 \vec{w}_1 + a_2 \vec{w}_2, \vec{w}_1 \rangle = a_1 \langle \vec{w}_1, \vec{w}_1 \rangle + a_2 \underbrace{\langle \vec{w}_2, \vec{w}_1 \rangle}_{=0} = a_1 \langle \vec{w}_1, \vec{w}_1 \rangle.$$

Therefore,

$$a_1 = \frac{\langle \vec{v}, \vec{w}_1 \rangle}{\langle \vec{w}_1, \vec{w}_1 \rangle}.$$

Similarly

$$a_2 = \frac{\langle \vec{v}, \vec{w}_2 \rangle}{\langle \vec{w}_2, \vec{w}_2 \rangle}.$$

You probably remember this formula from vector calculus.

Example 9.2.2: Write $\vec{v} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ as a linear combination of $\vec{w}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and $\vec{w}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

First note that \vec{w}_1 and \vec{w}_2 are orthogonal as $\langle \vec{w}_1, \vec{w}_2 \rangle = 1(1) + (-1)1 = 0$. Then

$$a_1 = \frac{\langle \vec{v}, \vec{w}_1 \rangle}{\langle \vec{w}_1, \vec{w}_1 \rangle} = \frac{2(1) + 3(-1)}{1(1) + (-1)(-1)} = \frac{-1}{2},$$

$$a_2 = \frac{\langle \vec{v}, \vec{w}_2 \rangle}{\langle \vec{w}_2, \vec{w}_2 \rangle} = \frac{2 + 3}{1 + 1} = \frac{5}{2}.$$

Hence

$$\begin{bmatrix} 2 \\ 3 \end{bmatrix} = \frac{-1}{2} \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \frac{5}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

9.2.3 The trigonometric series

Instead of decomposing a vector in terms of eigenvectors of a matrix, we decompose a function in terms of eigenfunctions of a certain eigenvalue problem. The eigenvalue problem we use for the Fourier series is

$$x'' + \lambda x = 0, \quad x(-\pi) = x(\pi), \quad x'(-\pi) = x'(\pi).$$

We computed that eigenfunctions are 1, $\cos(kt)$, $\sin(kt)$. That is, we want to find a representation of a 2π -periodic function $f(t)$ as

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nt) + b_n \sin(nt).$$

This series is called the *Fourier series** or the *trigonometric series* for $f(t)$. We write the coefficient of the eigenfunction 1 as $\frac{a_0}{2}$ for convenience. We could also think of 1 = $\cos(0t)$, so that we only need to look at $\cos(kt)$ and $\sin(kt)$.

As for matrices we want to find a *projection* of $f(t)$ onto the subspaces given by the eigenfunctions. So we want to define an *inner product of functions*. For example, to find a_n we want to compute $\langle f(t), \cos(nt) \rangle$. We define the inner product as

$$\langle f(t), g(t) \rangle \stackrel{\text{def}}{=} \int_{-\pi}^{\pi} f(t) g(t) dt.$$

With this definition of the inner product, we saw in the previous section that the eigenfunctions $\cos(kt)$ (including the constant eigenfunction), and $\sin(kt)$ are *orthogonal* in the sense that

$$\begin{aligned} \langle \cos(mt), \cos(nt) \rangle &= 0 && \text{for } m \neq n, \\ \langle \sin(mt), \sin(nt) \rangle &= 0 && \text{for } m \neq n, \\ \langle \sin(mt), \cos(nt) \rangle &= 0 && \text{for all } m \text{ and } n. \end{aligned}$$

For $n = 1, 2, 3, \dots$ we have

$$\begin{aligned} \langle \cos(nt), \cos(nt) \rangle &= \int_{-\pi}^{\pi} \cos(nt) \cos(nt) dt = \pi, \\ \langle \sin(nt), \sin(nt) \rangle &= \int_{-\pi}^{\pi} \sin(nt) \sin(nt) dt = \pi, \end{aligned}$$

by elementary calculus. For the constant we get

$$\langle 1, 1 \rangle = \int_{-\pi}^{\pi} 1 \cdot 1 dt = 2\pi.$$

The coefficients are given by

$$\begin{aligned} a_n &= \frac{\langle f(t), \cos(nt) \rangle}{\langle \cos(nt), \cos(nt) \rangle} = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \cos(nt) dt, \\ b_n &= \frac{\langle f(t), \sin(nt) \rangle}{\langle \sin(nt), \sin(nt) \rangle} = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \sin(nt) dt. \end{aligned}$$

*Named after the French mathematician [Jean Baptiste Joseph Fourier](#) (1768–1830).

Compare these expressions with the finite-dimensional example. For a_0 we get a similar formula

$$a_0 = 2 \frac{\langle f(t), 1 \rangle}{\langle 1, 1 \rangle} = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) dt.$$

Let us check the formulas using the orthogonality properties. Suppose for a moment that

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nt) + b_n \sin(nt).$$

Then for $m \geq 1$ we have

$$\begin{aligned} \langle f(t), \cos(mt) \rangle &= \left\langle \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nt) + b_n \sin(nt), \cos(mt) \right\rangle \\ &= \frac{a_0}{2} \langle 1, \cos(mt) \rangle + \sum_{n=1}^{\infty} a_n \langle \cos(nt), \cos(mt) \rangle + b_n \langle \sin(nt), \cos(mt) \rangle \\ &= a_m \langle \cos(mt), \cos(mt) \rangle. \end{aligned}$$

And hence $a_m = \frac{\langle f(t), \cos(mt) \rangle}{\langle \cos(mt), \cos(mt) \rangle}$.

Exercise 9.2.2: Carry out the calculation for a_0 and b_m .

Example 9.2.3: Take the function

$$f(t) = t$$

for t in $(-\pi, \pi]$. Extend $f(t)$ periodically and write it as a Fourier series. This function is called the *sawtooth*.

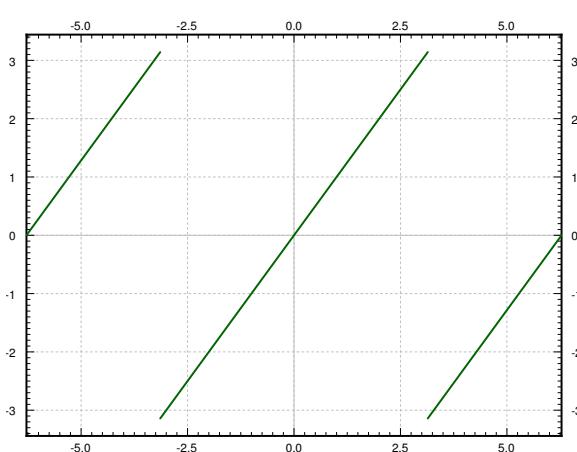


Figure 9.4: The graph of the sawtooth function.

Solution: The plot of the extended periodic function is given in Figure 9.4. Let us compute the coefficients. We start with a_0 ,

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} t dt = 0.$$

We will often use the result from calculus that says that the integral of an odd function over a symmetric interval is zero. Recall that an *odd function* is a function $\varphi(t)$ such that $\varphi(-t) = -\varphi(t)$. For example the functions t , $\sin t$, or (importantly for us) $t \cos(nt)$ are all odd functions. Thus

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} t \cos(nt) dt = 0.$$

Let us move to b_n . Another useful fact from calculus is that the integral of an even function over a symmetric interval is twice the integral of the same function over half the interval. Recall an *even function* is a function $\varphi(t)$ such that $\varphi(-t) = \varphi(t)$. For example $t \sin(nt)$ is even.

$$\begin{aligned} b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} t \sin(nt) dt \\ &= \frac{2}{\pi} \int_0^{\pi} t \sin(nt) dt \\ &= \frac{2}{\pi} \left(\left[\frac{-t \cos(nt)}{n} \right]_{t=0}^{\pi} + \frac{1}{n} \int_0^{\pi} \cos(nt) dt \right) \\ &= \frac{2}{\pi} \left(\frac{-\pi \cos(n\pi)}{n} + 0 \right) \\ &= \frac{-2 \cos(n\pi)}{n} = \frac{2(-1)^{n+1}}{n}. \end{aligned}$$

We have used the fact that

$$\cos(n\pi) = (-1)^n = \begin{cases} 1 & \text{if } n \text{ even,} \\ -1 & \text{if } n \text{ odd.} \end{cases}$$

The series, therefore, is

$$\sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin(nt).$$

Let us write out the first 3 harmonics of the series for $f(t)$.

$$2 \sin(t) - \sin(2t) + \frac{2}{3} \sin(3t) + \dots$$

The plot of these first three terms of the series, along with a plot of the first 20 terms is given in [Figure 9.5](#) on the facing page. \(\square\)

Example 9.2.4: Take the function

$$f(t) = \begin{cases} 0 & \text{if } -\pi < t \leq 0, \\ \pi & \text{if } 0 < t \leq \pi. \end{cases}$$

Extend $f(t)$ periodically and write it as a Fourier series. This function or its variants appear often in applications and the function is called the *square wave*.

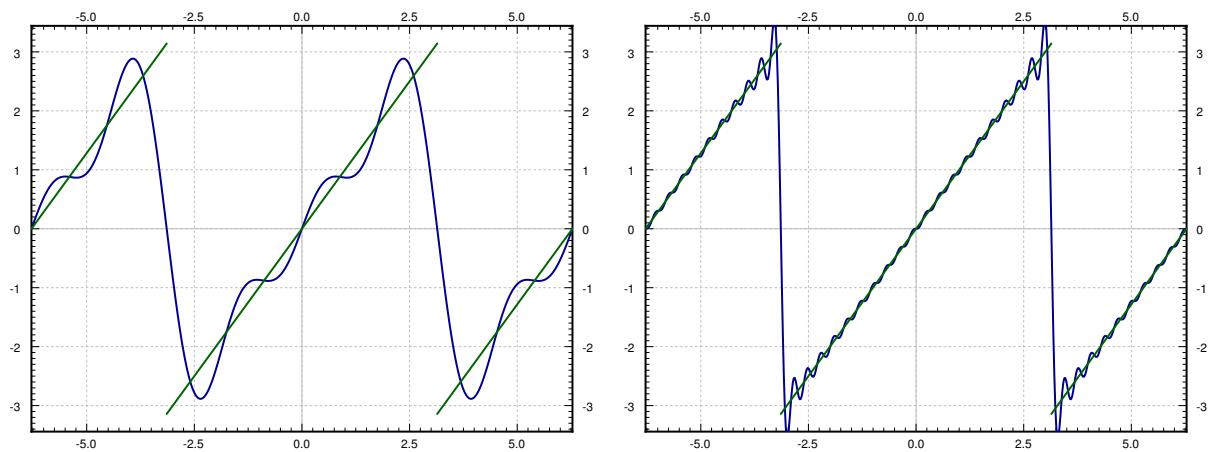


Figure 9.5: First 3 (left graph) and 20 (right graph) harmonics of the sawtooth function.

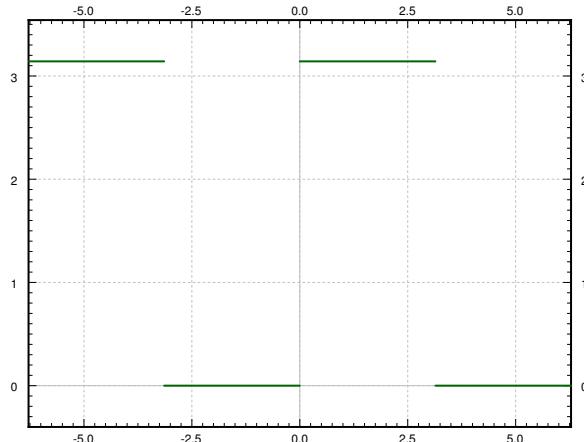


Figure 9.6: The graph of the square wave function.

Solution: The plot of the extended periodic function is given in Figure 9.6. Now we compute the coefficients. Let us start with a_0

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) dt = \frac{1}{\pi} \int_0^{\pi} \pi dt = \pi.$$

Next,

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \cos(nt) dt = \frac{1}{\pi} \int_0^{\pi} \pi \cos(nt) dt = 0.$$

And finally

$$\begin{aligned}
 b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \sin(nt) dt \\
 &= \frac{1}{\pi} \int_0^\pi \pi \sin(nt) dt \\
 &= \left[\frac{-\cos(nt)}{n} \right]_{t=0}^\pi \\
 &= \frac{1 - \cos(\pi n)}{n} = \frac{1 - (-1)^n}{n} = \begin{cases} \frac{2}{n} & \text{if } n \text{ is odd,} \\ 0 & \text{if } n \text{ is even.} \end{cases}
 \end{aligned}$$

The Fourier series is

$$\frac{\pi}{2} + \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{2}{n} \sin(nt) = \frac{\pi}{2} + \sum_{k=1}^{\infty} \frac{2}{2k-1} \sin((2k-1)t).$$

Let us write out the first 3 harmonics of the series for $f(t)$.

$$\frac{\pi}{2} + 2 \sin(t) + \frac{2}{3} \sin(3t) + \dots$$

The plot of these first three and also of the first 20 terms of the series is given in [Figure 9.7](#).

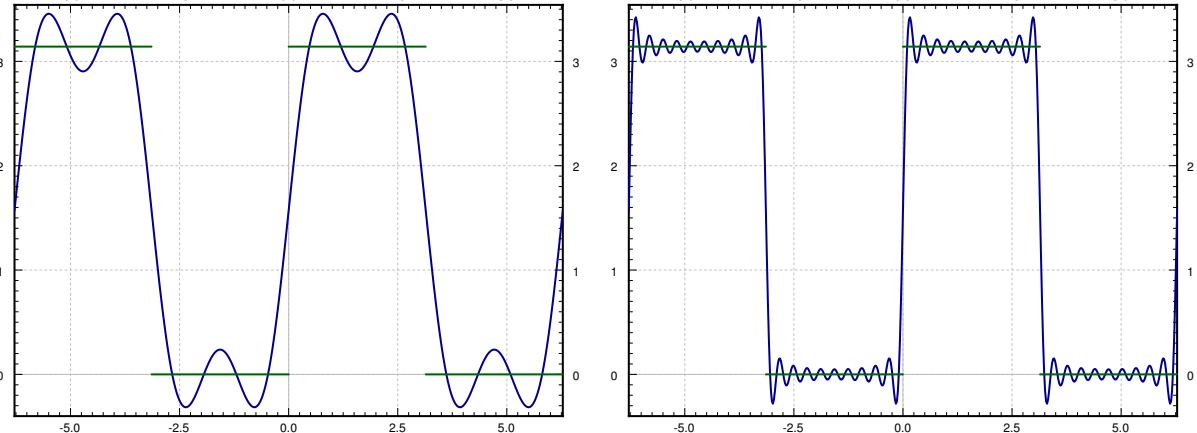


Figure 9.7: First 3 (left graph) and 20 (right graph) harmonics of the square wave function.

We have so far skirted the issue of convergence. For example, if $f(t)$ is the square wave function, the equation

$$f(t) = \frac{\pi}{2} + \sum_{k=1}^{\infty} \frac{2}{2k-1} \sin((2k-1)t).$$

is only an equality for such t where $f(t)$ is continuous. That is, we do not get an equality for $t = -\pi, 0, \pi$ and all the other discontinuities of $f(t)$. It is not hard to see that when t is an integer multiple of π (which includes all the discontinuities), then

$$\frac{\pi}{2} + \sum_{k=1}^{\infty} \frac{2}{2k-1} \sin((2k-1)t) = \frac{\pi}{2}.$$

We redefine $f(t)$ on $[-\pi, \pi]$ as

$$f(t) = \begin{cases} 0 & \text{if } -\pi < t < 0, \\ \pi & \text{if } 0 < t < \pi, \\ \pi/2 & \text{if } t = -\pi, t = 0, \text{ or } t = \pi, \end{cases}$$

and extend periodically. The series equals this extended $f(t)$ everywhere, including the discontinuities. We will generally not worry about changing the function values at several (finitely many) points.

We will say more about convergence in the next section. Let us however mention briefly an effect of the discontinuity. Let us zoom in near the discontinuity in the square wave. Further, let us plot the first 100 harmonics, see [Figure 9.8](#). While the series is a very good approximation away from the discontinuities, the error (the overshoot) near the discontinuity at $t = \pi$ does not seem to be getting any smaller. This behavior is known as the *Gibbs phenomenon*. The region where the error is large does get smaller, however, the more terms in the series we take.

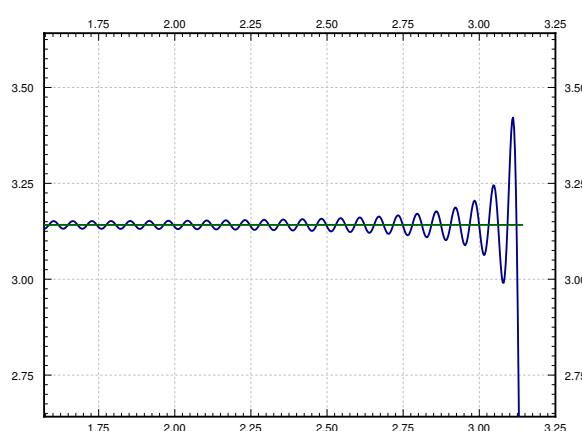


Figure 9.8: Gibbs phenomenon in action.

We can think of a periodic function as a “signal” being a superposition of many signals of pure frequency. For example, we could think of the square wave as a tone of certain base frequency. This base frequency is called the *fundamental frequency*. The square wave will be a superposition of many different pure tones of frequencies that are multiples of the fundamental frequency. In music, the higher frequencies are called the *overtones*. All the frequencies that appear are called the *spectrum* of the signal. On the other hand a simple

sine wave is only the pure tone (no overtones). The simplest way to make sound using a computer is the square wave, and the sound is very different from a pure tone. If you ever played video games from the 1980s or so, then you heard what square waves sound like.

9.2.4 Exercises

Exercise 9.2.3:* Suppose $f(t)$ is defined on $[-\pi, \pi]$ as $f(t) = \sin(t)$. Extend periodically and compute the Fourier series.

Exercise 9.2.4: Suppose $f(t)$ is defined on $[-\pi, \pi]$ as $\sin(5t) + \cos(3t)$. Extend periodically and compute the Fourier series of $f(t)$.

Exercise 9.2.5: Suppose $f(t)$ is defined on $[-\pi, \pi]$ as $|t|$. Extend periodically and compute the Fourier series of $f(t)$.

Exercise 9.2.6: Suppose $f(t)$ is defined on $[-\pi, \pi]$ as $|t|^3$. Extend periodically and compute the Fourier series of $f(t)$.

Exercise 9.2.7:* Suppose $f(t)$ is defined on $(-\pi, \pi]$ as $f(t) = \sin(\pi t)$. Extend periodically and compute the Fourier series.

Exercise 9.2.8: Suppose $f(t)$ is defined on $(-\pi, \pi]$ as

$$f(t) = \begin{cases} -1 & \text{if } -\pi < t \leq 0, \\ 1 & \text{if } 0 < t \leq \pi. \end{cases}$$

Extend periodically and compute the Fourier series of $f(t)$.

Exercise 9.2.9: Suppose $f(t)$ is defined on $(-\pi, \pi]$ as t^3 . Extend periodically and compute the Fourier series of $f(t)$.

Exercise 9.2.10:* Suppose $f(t)$ is defined on $(-\pi, \pi]$ as $f(t) = \sin^2(t)$. Extend periodically and compute the Fourier series.

Exercise 9.2.11: Suppose $f(t)$ is defined on $[-\pi, \pi]$ as t^2 . Extend periodically and compute the Fourier series of $f(t)$.

Exercise 9.2.12:* Suppose $f(t)$ is defined on $(-\pi, \pi]$ as $f(t) = t^4$. Extend periodically and compute the Fourier series.

There is another form of the Fourier series using complex exponentials e^{nt} for $n = \dots, -2, -1, 0, 1, 2, \dots$ instead of $\cos(nt)$ and $\sin(nt)$ for positive n . This form may be easier to work with sometimes. It is certainly more compact to write, and there is only one formula for the coefficients. On the downside, the coefficients are complex numbers.

Exercise 9.2.13: Let

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nt) + b_n \sin(nt).$$

Use Euler's formula $e^{i\theta} = \cos(\theta) + i \sin(\theta)$ to show that there exist complex numbers c_m such that

$$f(t) = \sum_{m=-\infty}^{\infty} c_m e^{imt}.$$

Note that the sum now ranges over all the integers including negative ones. Do not worry about convergence in this calculation. Hint: It may be better to start from the complex exponential form and write the series as

$$c_0 + \sum_{m=1}^{\infty} \left(c_m e^{imt} + c_{-m} e^{-imt} \right).$$

9.3 More on the Fourier series

Attribution: [JL], §4.3.

Learning Objectives

After this section, you will be able to:

- Discuss Fourier series over intervals of different lengths,
- Discuss the convergence of Fourier series, and
- Compute derivatives and integrals of functions written as Fourier series.

9.3.1 $2L$ -periodic functions

We have computed the Fourier series for a 2π -periodic function, but what about functions of different periods. Well, fear not, the computation is a simple case of change of variables. We just rescale the independent axis. Suppose we have a $2L$ -periodic function $f(t)$. Then L is called the *half period*. Let $s = \frac{\pi}{L}t$. Then the function

$$g(s) = f\left(\frac{L}{\pi}s\right)$$

is 2π -periodic. We must also rescale all our sines and cosines. In the series we use $\frac{\pi}{L}t$ as the variable. That is, we want to write

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}t\right) + b_n \sin\left(\frac{n\pi}{L}t\right).$$

If we change variables to s , we see that

$$g(s) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(ns) + b_n \sin(ns).$$

We compute a_n and b_n as before. After we write down the integrals, we change variables from s back to t , noting also that $ds = \frac{\pi}{L} dt$.

$$\begin{aligned} a_0 &= \frac{1}{\pi} \int_{-\pi}^{\pi} g(s) ds = \frac{1}{L} \int_{-L}^L f(t) dt, \\ a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} g(s) \cos(ns) ds = \frac{1}{L} \int_{-L}^L f(t) \cos\left(\frac{n\pi}{L}t\right) dt, \\ b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} g(s) \sin(ns) ds = \frac{1}{L} \int_{-L}^L f(t) \sin\left(\frac{n\pi}{L}t\right) dt. \end{aligned}$$

The two most common half periods that show up in examples are π and 1 because of the simplicity of the formulas. We should stress that we have done no new mathematics, we have only changed variables. If you understand the Fourier series for 2π -periodic functions, you understand it for $2L$ -periodic functions. You can think of it as just using different units for time. All that we are doing is moving some constants around, but all the mathematics is the same.

Example 9.3.1: Let

$$f(t) = |t| \quad \text{for } -1 < t \leq 1,$$

extended periodically. The plot of the periodic extension is given in Figure 9.9. Compute the Fourier series of $f(t)$.

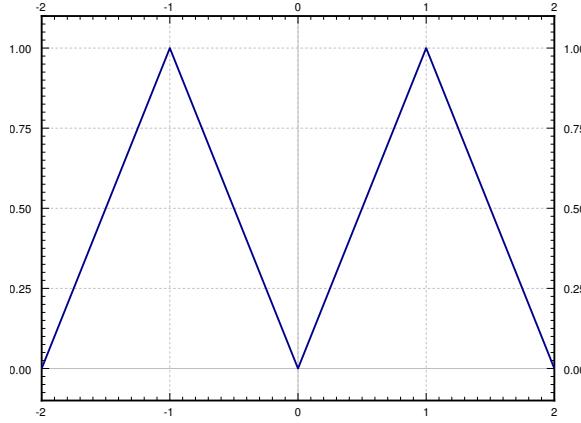


Figure 9.9: Periodic extension of the function $f(t)$.

Solution: We want to write $f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\pi t) + b_n \sin(n\pi t)$. For $n \geq 1$ we note that $|t| \cos(n\pi t)$ is even and hence

$$\begin{aligned} a_n &= \int_{-1}^1 f(t) \cos(n\pi t) dt \\ &= 2 \int_0^1 t \cos(n\pi t) dt \\ &= 2 \left[\frac{t}{n\pi} \sin(n\pi t) \right]_{t=0}^1 - 2 \int_0^1 \frac{1}{n\pi} \sin(n\pi t) dt \\ &= 0 + \frac{1}{n^2\pi^2} \left[\cos(n\pi t) \right]_{t=0}^1 = \frac{2((-1)^n - 1)}{n^2\pi^2} = \begin{cases} 0 & \text{if } n \text{ is even,} \\ \frac{-4}{n^2\pi^2} & \text{if } n \text{ is odd.} \end{cases} \end{aligned}$$

Next we find a_0 :

$$a_0 = \int_{-1}^1 |t| dt = 1.$$

You should be able to find this integral by thinking about the integral as the area under the graph without doing any computation at all. Finally we can find b_n . Here, we notice that $|t| \sin(n\pi t)$ is odd and, therefore,

$$b_n = \int_{-1}^1 f(t) \sin(n\pi t) dt = 0.$$

Hence, the series is

$$\frac{1}{2} + \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{-4}{n^2 \pi^2} \cos(n\pi t).$$

Let us explicitly write down the first few terms of the series up to the 3rd harmonic.

$$\frac{1}{2} - \frac{4}{\pi^2} \cos(\pi t) - \frac{4}{9\pi^2} \cos(3\pi t) - \dots$$

The plot of these few terms and also a plot up to the 20th harmonic is given in [Figure 9.10](#). You should notice how close the graph is to the real function. You should also notice that there is no “Gibbs phenomenon” present as there are no discontinuities.

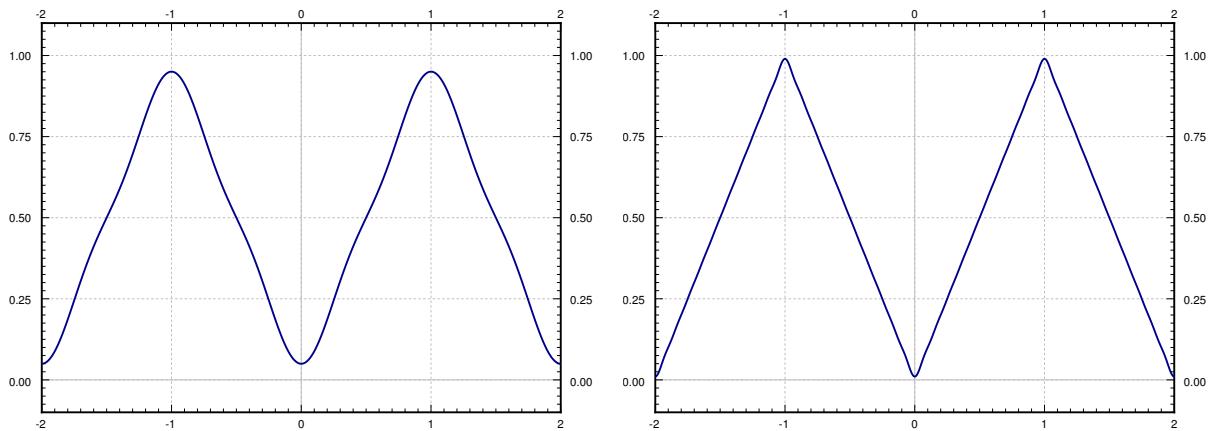


Figure 9.10: Fourier series of $f(t)$ up to the 3rd harmonic (left graph) and up to the 20th harmonic (right graph).

9.3.2 Convergence

We will need the one sided limits of functions. We will use the following notation

$$f(c-) = \lim_{t \uparrow c} f(t), \quad \text{and} \quad f(c+) = \lim_{t \downarrow c} f(t).$$

If you are unfamiliar with this notation, $\lim_{t \uparrow c} f(t)$ means we are taking a limit of $f(t)$ as t approaches c from below (i.e. $t < c$) and $\lim_{t \downarrow c} f(t)$ means we are taking a limit of $f(t)$ as t

approaches c from above (i.e. $t > c$). For example, for the square wave function

$$f(t) = \begin{cases} 0 & \text{if } -\pi < t \leq 0, \\ \pi & \text{if } 0 < t \leq \pi, \end{cases} \quad (9.8)$$

we have $f(0-) = 0$ and $f(0+) = \pi$.

Let $f(t)$ be a function defined on an interval $[a, b]$. Suppose that we find finitely many points $a = t_0, t_1, t_2, \dots, t_k = b$ in the interval, such that $f(t)$ is continuous on the intervals $(t_0, t_1), (t_1, t_2), \dots, (t_{k-1}, t_k)$. Also suppose that all the one sided limits exist, that is, all of $f(t_0+), f(t_1-), f(t_1+), f(t_2-), \dots, f(t_k-)$ exist and are finite. Then we say $f(t)$ is *piecewise continuous*.

If moreover, $f(t)$ is differentiable at all but finitely many points, and $f'(t)$ is piecewise continuous, then $f(t)$ is said to be *piecewise smooth*.

Example 9.3.2: The square wave function (9.8) is piecewise smooth on $[-\pi, \pi]$ or any other interval. In such a case we simply say that the function is piecewise smooth.

Example 9.3.3: The function $f(t) = |t|$ is piecewise smooth.

Example 9.3.4: The function $f(t) = \frac{1}{t}$ is not piecewise smooth on $[-1, 1]$ (or any other interval containing zero). In fact, it is not even piecewise continuous.

Example 9.3.5: The function $f(t) = \sqrt[3]{t}$ is not piecewise smooth on $[-1, 1]$ (or any other interval containing zero). $f(t)$ is continuous, but the derivative of $f(t)$ is unbounded near zero and hence not piecewise continuous.

Piecewise smooth functions have an easy answer on the convergence of the Fourier series.

Theorem 9.3.1

Suppose $f(t)$ is a $2L$ -periodic piecewise smooth function. Let

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}t\right) + b_n \sin\left(\frac{n\pi}{L}t\right)$$

be the Fourier series for $f(t)$. Then the series converges for all t . If $f(t)$ is continuous at t , then

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}t\right) + b_n \sin\left(\frac{n\pi}{L}t\right).$$

Otherwise,

$$\frac{f(t-) + f(t+)}{2} = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}t\right) + b_n \sin\left(\frac{n\pi}{L}t\right).$$

If we happen to have that $f(t) = \frac{f(t-) + f(t+)}{2}$ at all the discontinuities, the Fourier series converges to $f(t)$ everywhere. We can always just redefine $f(t)$ by changing the value at each discontinuity appropriately. Then we can write an equals sign between $f(t)$ and the series without any worry. We mentioned this fact briefly at the end last section.

The theorem does not say how fast the series converges. Think back to the discussion of the Gibbs phenomenon in the last section. The closer you get to the discontinuity, the more terms you need to take to get an accurate approximation to the function.

9.3.3 Differentiation and integration of Fourier series

Not only does Fourier series converge nicely, but it is easy to differentiate and integrate the series. We can do this just by differentiating or integrating term by term.

Theorem 9.3.2

Suppose

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}t\right) + b_n \sin\left(\frac{n\pi}{L}t\right)$$

is a piecewise smooth continuous function and the derivative $f'(t)$ is piecewise smooth. Then the derivative can be obtained by differentiating term by term,

$$f'(t) = \sum_{n=1}^{\infty} \frac{-a_n n\pi}{L} \sin\left(\frac{n\pi}{L}t\right) + \frac{b_n n\pi}{L} \cos\left(\frac{n\pi}{L}t\right).$$

It is important that the function is continuous. It can have corners, but no jumps. Otherwise, the differentiated series will fail to converge. For an exercise, take the series obtained for the square wave and try to differentiate the series. Similarly, we can also integrate a Fourier series.

Theorem 9.3.3

Suppose

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}t\right) + b_n \sin\left(\frac{n\pi}{L}t\right)$$

is a piecewise smooth function. Then the antiderivative is obtained by antidifferentiating term by term and so

$$F(t) = \frac{a_0 t}{2} + C + \sum_{n=1}^{\infty} \frac{a_n L}{n\pi} \sin\left(\frac{n\pi}{L}t\right) + \frac{-b_n L}{n\pi} \cos\left(\frac{n\pi}{L}t\right),$$

where $F'(t) = f(t)$ and C is an arbitrary constant.

Note that the series for $F(t)$ is no longer a Fourier series as it contains the $\frac{a_0 t}{2}$ term. The antiderivative of a periodic function need no longer be periodic and so we should not expect a Fourier series.

9.3.4 Rates of convergence and smoothness

Let us do an example of a periodic function with one derivative everywhere.

Example 9.3.6: Take the function

$$f(t) = \begin{cases} (t+1)t & \text{if } -1 < t \leq 0, \\ (1-t)t & \text{if } 0 < t \leq 1, \end{cases}$$

and extend to a 2-periodic function. The plot is given in Figure 9.11.

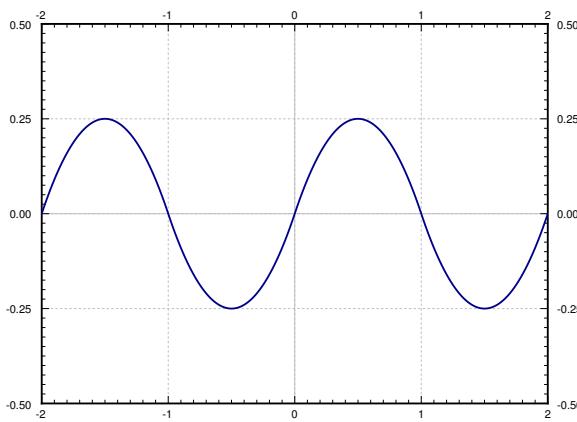


Figure 9.11: Smooth 2-periodic function.

This function has one derivative everywhere, but it does not have a second derivative whenever t is an integer.

Exercise 9.3.1: Compute $f''(0+)$ and $f''(0-)$.

Let us compute the Fourier series coefficients. The actual computation involves several integration by parts and is left to student.

$$\begin{aligned} a_0 &= \int_{-1}^1 f(t) dt = \int_{-1}^0 (t+1)t dt + \int_0^1 (1-t)t dt = 0, \\ a_n &= \int_{-1}^1 f(t) \cos(n\pi t) dt = \int_{-1}^0 (t+1)t \cos(n\pi t) dt + \int_0^1 (1-t)t \cos(n\pi t) dt = 0, \\ b_n &= \int_{-1}^1 f(t) \sin(n\pi t) dt = \int_{-1}^0 (t+1)t \sin(n\pi t) dt + \int_0^1 (1-t)t \sin(n\pi t) dt \\ &= \frac{4(1 - (-1)^n)}{\pi^3 n^3} = \begin{cases} \frac{8}{\pi^3 n^3} & \text{if } n \text{ is odd,} \\ 0 & \text{if } n \text{ is even.} \end{cases} \end{aligned}$$

That is, the series is

$$\sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{8}{\pi^3 n^3} \sin(n\pi t).$$

This series converges very fast. If you plot up to the third harmonic, that is the function

$$\frac{8}{\pi^3} \sin(\pi t) + \frac{8}{27\pi^3} \sin(3\pi t),$$

it is almost indistinguishable from the plot of $f(t)$ in [Figure 9.11](#) on the previous page. In fact, the coefficient $\frac{8}{27\pi^3}$ is already just 0.0096 (approximately). The reason for this behavior is the n^3 term in the denominator. The coefficients b_n in this case go to zero as fast as $1/n^3$ goes to zero.

For functions constructed piecewise from polynomials as above, it is generally true that if you have one derivative, the Fourier coefficients will go to zero approximately like $1/n^3$. If you have only a continuous function, then the Fourier coefficients will go to zero as $1/n^2$. If you have discontinuities, then the Fourier coefficients will go to zero approximately as $1/n$. For more general functions the story is somewhat more complicated but the same idea holds, the more derivatives you have, the faster the coefficients go to zero. Similar reasoning works in reverse. If the coefficients go to zero like $1/n^2$, you always obtain a continuous function. If they go to zero like $1/n^3$, you obtain an everywhere differentiable function.

To justify this behavior, take for example the function defined by the Fourier series

$$f(t) = \sum_{n=1}^{\infty} \frac{1}{n^3} \sin(nt).$$

When we differentiate term by term we notice

$$f'(t) = \sum_{n=1}^{\infty} \frac{1}{n^2} \cos(nt).$$

Therefore, the coefficients now go down like $1/n^2$, which means that we have a continuous function. The derivative of $f'(t)$ is defined at most points, but there are points where $f'(t)$ is not differentiable. It has corners, but no jumps. If we differentiate again (where we can), we find that the function $f''(t)$, now fails to be continuous (has jumps)

$$f''(t) = \sum_{n=1}^{\infty} \frac{-1}{n} \sin(nt).$$

This function is similar to the sawtooth. If we tried to differentiate the series again, we would obtain

$$\sum_{n=1}^{\infty} -\cos(nt),$$

which does not converge!

Exercise 9.3.2: Use a computer to plot the series we obtained for $f(t)$, $f'(t)$ and $f''(t)$. That is, plot say the first 5 harmonics of the functions. At what points does $f''(t)$ have the discontinuities?

9.3.5 Exercises

Exercise 9.3.3: Let

$$f(t) = \begin{cases} 0 & \text{if } -1 < t \leq 0, \\ t & \text{if } 0 < t \leq 1, \end{cases}$$

extended periodically.

- a) Compute the Fourier series for $f(t)$.
 b) Write out the series explicitly up to the 3rd harmonic.

Exercise 9.3.4:* Let

$$f(t) = t^2 \quad \text{for } -2 < t \leq 2$$

extended periodically.

- a) Compute the Fourier series for $f(t)$.
 b) Write out the series explicitly up to the 3rd harmonic.

Exercise 9.3.5: Let

$$f(t) = \begin{cases} -t & \text{if } -1 < t \leq 0, \\ t^2 & \text{if } 0 < t \leq 1, \end{cases}$$

extended periodically.

- a) Compute the Fourier series for $f(t)$.
 b) Write out the series explicitly up to the 3rd harmonic.

Exercise 9.3.6:* Let

$$f(t) = t \quad \text{for } -\lambda < t \leq \lambda \quad (\text{for some } \lambda > 0)$$

extended periodically.

- a) Compute the Fourier series for $f(t)$.
 b) Write out the series explicitly up to the 3rd harmonic.

Exercise 9.3.7: Let

$$f(t) = \begin{cases} \frac{-t}{10} & \text{if } -10 < t \leq 0, \\ \frac{t}{10} & \text{if } 0 < t \leq 10, \end{cases}$$

extended periodically (period is 20).

- a) Compute the Fourier series for $f(t)$.
 b) Write out the series explicitly up to the 3rd harmonic.

Exercise 9.3.8:* Let

$$f(t) = \frac{1}{2} + \sum_{n=1}^{\infty} \frac{1}{n(n^2+1)} \sin(n\pi t).$$

Compute $f'(t)$.

Exercise 9.3.9: Let $f(t) = \sum_{n=1}^{\infty} \frac{1}{n^3} \cos(nt)$. Is $f(t)$ continuous and differentiable everywhere? Find the derivative (if it exists everywhere) or justify why $f(t)$ is not differentiable everywhere.

Exercise 9.3.10: Let $f(t) = \sum_{n=1}^{\infty} \frac{(-1)^n}{n} \sin(nt)$. Is $f(t)$ differentiable everywhere? Find the derivative (if it exists everywhere) or justify why $f(t)$ is not differentiable everywhere.

Exercise 9.3.11:* Let

$$f(t) = \frac{1}{2} + \sum_{n=1}^{\infty} \frac{1}{n^3} \cos(nt).$$

- a) Find the antiderivative.
- b) Is the antiderivative periodic?

Exercise 9.3.12: Let

$$f(t) = \begin{cases} 0 & \text{if } -2 < t \leq 0, \\ t & \text{if } 0 < t \leq 1, \\ -t + 2 & \text{if } 1 < t \leq 2, \end{cases}$$

extended periodically.

- a) Compute the Fourier series for $f(t)$.
- b) Write out the series explicitly up to the 3rd harmonic.

Exercise 9.3.13: Let

$$f(t) = e^t \quad \text{for } -1 < t \leq 1$$

extended periodically.

- a) Compute the Fourier series for $f(t)$.
- b) Write out the series explicitly up to the 3rd harmonic.
- c) What does the series converge to at $t = 1$.

Exercise 9.3.14: Let

$$f(t) = t^2 \quad \text{for } -1 < t \leq 1$$

extended periodically.

- a) Compute the Fourier series for $f(t)$.
- b) By plugging in $t = 0$, evaluate $\sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} = 1 - \frac{1}{4} + \frac{1}{9} - \dots$
- c) Now evaluate $\sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \frac{1}{4} + \frac{1}{9} + \dots$

Exercise 9.3.15:* Let

$$f(t) = t/2 \quad \text{for } -\pi < t < \pi$$

extended periodically.

- a) Compute the Fourier series for $f(t)$.
- b) Plug in $t = \pi/2$ to find a series representation for $\pi/4$.
- c) Using the first 4 terms of the result from part b) approximate $\pi/4$.

Exercise 9.3.16: Let

$$f(t) = \begin{cases} 0 & \text{if } -3 < t \leq 0, \\ t & \text{if } 0 < t \leq 3, \end{cases}$$

extended periodically. Suppose $F(t)$ is the function given by the Fourier series of f . Without computing the Fourier series evaluate

- a) $F(2)$
- b) $F(-2)$
- c) $F(4)$
- d) $F(-4)$
- e) $F(3)$
- f) $F(-9)$

Exercise 9.3.17:* Let

$$f(t) = \begin{cases} 0 & \text{if } -2 < t \leq 0, \\ 2 & \text{if } 0 < t \leq 2, \end{cases}$$

extended periodically. Suppose $F(t)$ is the function given by the Fourier series of f . Without computing the Fourier series evaluate

- a) $F(0)$
- b) $F(-1)$
- c) $F(1)$
- d) $F(-2)$
- e) $F(4)$
- f) $F(-8)$

9.4 Sine and cosine series

Attribution: [JL], §4.4.

Learning Objectives

After this section, you will be able to:

- Use sine and cosine series to represent odd and even periodic extensions of functions and
- Understand the connection between Fourier series and boundary value problems.

9.4.1 Odd and even periodic functions

You may have noticed by now that an odd function has no cosine terms in the Fourier series and an even function has no sine terms in the Fourier series. This observation is not a coincidence. Let us look at even and odd periodic function in more detail.

Recall that a function $f(t)$ is *odd* if $f(-t) = -f(t)$. A function $f(t)$ is *even* if $f(-t) = f(t)$. For example, $\cos(nt)$ is even and $\sin(nt)$ is odd. Similarly the function t^k is even if k is even and odd when k is odd.

Exercise 9.4.1: Take two functions $f(t)$ and $g(t)$ and define their product $h(t) = f(t)g(t)$.

- Suppose both $f(t)$ and $g(t)$ are odd. Is $h(t)$ odd or even?
- Suppose one is even and one is odd. Is $h(t)$ odd or even?
- Suppose both are even. Is $h(t)$ odd or even?

If $f(t)$ and $g(t)$ are both odd, then $f(t) + g(t)$ is odd. Similarly for even functions. On the other hand, if $f(t)$ is odd and $g(t)$ even, then we cannot say anything about the sum $f(t) + g(t)$. In fact, the Fourier series of any function is a sum of an odd (the sine terms) and an even (the cosine terms) function.

In this section we consider odd and even periodic functions. We have previously defined the $2L$ -periodic extension of a function defined on the interval $[-L, L]$. Sometimes we are only interested in the function on the range $[0, L]$ and it would be convenient to have an odd (resp. even) function. If the function is odd (resp. even), all the cosine (resp. sine) terms disappear. What we will do is take the odd (resp. even) extension of the function to $[-L, L]$ and then extend periodically to a $2L$ -periodic function.

Take a function $f(t)$ defined on $[0, L]$. On $(-L, L]$ define the functions

$$\begin{aligned} F_{\text{odd}}(t) &\stackrel{\text{def}}{=} \begin{cases} f(t) & \text{if } 0 \leq t \leq L, \\ -f(-t) & \text{if } -L < t < 0, \end{cases} \\ F_{\text{even}}(t) &\stackrel{\text{def}}{=} \begin{cases} f(t) & \text{if } 0 \leq t \leq L, \\ f(-t) & \text{if } -L < t < 0. \end{cases} \end{aligned}$$

Extend $F_{\text{odd}}(t)$ and $F_{\text{even}}(t)$ to be $2L$ -periodic. Then $F_{\text{odd}}(t)$ is called the *odd periodic extension* of $f(t)$, and $F_{\text{even}}(t)$ is called the *even periodic extension* of $f(t)$. For the odd extension we generally assume that $f(0) = f(L) = 0$.

Exercise 9.4.2: Check that $F_{\text{odd}}(t)$ is odd and $F_{\text{even}}(t)$ is even. For F_{odd} , assume $f(0) = f(L) = 0$.

Example 9.4.1: Take the function $f(t) = t(1-t)$ defined on $[0, 1]$. Figure 9.12 shows the plots of the odd and even periodic extensions of $f(t)$.

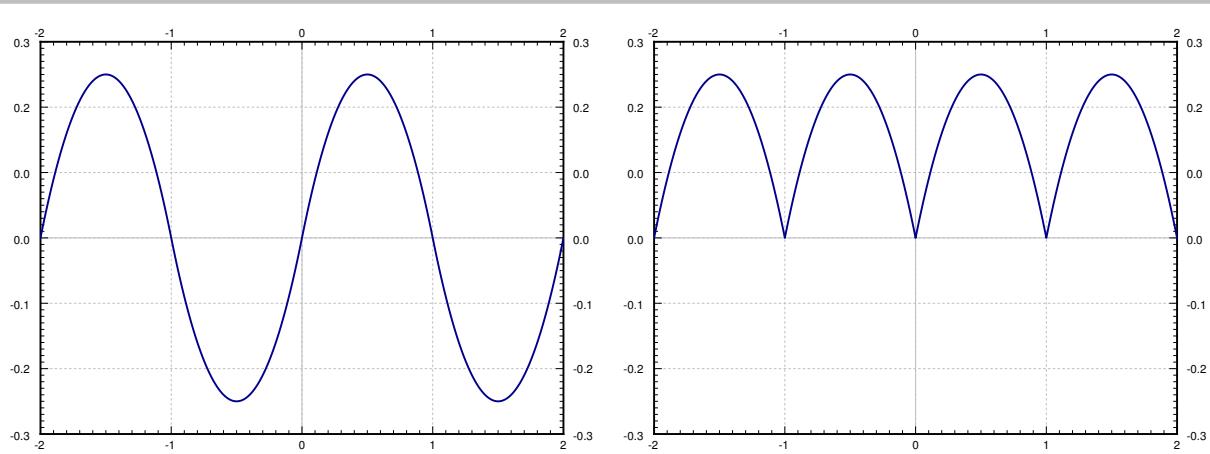


Figure 9.12: Odd and even 2-periodic extension of $f(t) = t(1-t)$, $0 \leq t \leq 1$.

9.4.2 Sine and cosine series

Let $f(t)$ be an odd $2L$ -periodic function. We write the Fourier series for $f(t)$. First, we compute the coefficients a_n (including $n = 0$) and get

$$a_n = \frac{1}{L} \int_{-L}^L f(t) \cos\left(\frac{n\pi}{L}t\right) dt = 0.$$

That is, there are no cosine terms in the Fourier series of an odd function. The integral is zero because $f(t) \cos(n\pi Lt)$ is an odd function (product of an odd and an even function is odd) and the integral of an odd function over a symmetric interval is always zero. The integral of an even function over a symmetric interval $[-L, L]$ is twice the integral of the function over the interval $[0, L]$. The function $f(t) \sin\left(\frac{n\pi}{L}t\right)$ is the product of two odd functions and hence is even.

$$b_n = \frac{1}{L} \int_{-L}^L f(t) \sin\left(\frac{n\pi}{L}t\right) dt = \frac{2}{L} \int_0^L f(t) \sin\left(\frac{n\pi}{L}t\right) dt.$$

We now write the Fourier series of $f(t)$ as

$$\sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi}{L}t\right).$$

Similarly, if $f(t)$ is an even $2L$ -periodic function. For the same exact reasons as above, we find that $b_n = 0$ and

$$a_n = \frac{2}{L} \int_0^L f(t) \cos\left(\frac{n\pi}{L}t\right) dt.$$

The formula still works for $n = 0$, in which case it becomes

$$a_0 = \frac{2}{L} \int_0^L f(t) dt.$$

The Fourier series is then

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}t\right).$$

An interesting consequence is that the coefficients of the Fourier series of an odd (or even) function can be computed by just integrating over the half interval $[0, L]$. Therefore, we can compute the Fourier series of the odd (or even) extension of a function by computing certain integrals over the interval where the original function is defined.

Theorem 9.4.1

Let $f(t)$ be a piecewise smooth function defined on $[0, L]$. Then the odd periodic extension of $f(t)$ has the Fourier series

$$F_{\text{odd}}(t) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi}{L}t\right),$$

where

$$b_n = \frac{2}{L} \int_0^L f(t) \sin\left(\frac{n\pi}{L}t\right) dt.$$

The even periodic extension of $f(t)$ has the Fourier series

$$F_{\text{even}}(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}t\right),$$

where

$$a_n = \frac{2}{L} \int_0^L f(t) \cos\left(\frac{n\pi}{L}t\right) dt.$$

Definition 9.4.1

We call the series $\sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi}{L}t\right)$ the *sine series* of $f(t)$ and we call the series $\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}t\right)$ the *cosine series* of $f(t)$.

We often do not actually care what happens outside of $[0, L]$. In this case, we pick whichever series fits our problem better.

It is not necessary to start with the full Fourier series to obtain the sine and cosine series. The sine series is really the eigenfunction expansion of $f(t)$ using eigenfunctions of the

eigenvalue problem $x'' + \lambda x = 0$, $x(0) = 0$, $x(L) = L$. The cosine series is the eigenfunction expansion of $f(t)$ using eigenfunctions of the eigenvalue problem $x'' + \lambda x = 0$, $x'(0) = 0$, $x'(L) = L$. We could have, therefore, gotten the same formulas by defining the inner product

$$\langle f(t), g(t) \rangle = \int_0^L f(t)g(t) dt,$$

and following the procedure of § 9.2. This point of view is useful, as we commonly use a specific series that arose because our underlying question led to a certain eigenvalue problem. If the eigenvalue problem is not one of the three we covered so far, you can still do an eigenfunction expansion, generalizing the results of this chapter.

Example 9.4.2: Find the Fourier series of the even periodic extension of the function $f(t) = t^2$ for $0 \leq t \leq \pi$.

Solution: We want to write

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nt),$$

where

$$a_0 = \frac{2}{\pi} \int_0^\pi t^2 dt = \frac{2\pi^2}{3},$$

and

$$\begin{aligned} a_n &= \frac{2}{\pi} \int_0^\pi t^2 \cos(nt) dt = \frac{2}{\pi} \left[t^2 \frac{1}{n} \sin(nt) \right]_0^\pi - \frac{4}{n\pi} \int_0^\pi t \sin(nt) dt \\ &= \frac{4}{n^2\pi} \left[t \cos(nt) \right]_0^\pi + \frac{4}{n^2\pi} \int_0^\pi \cos(nt) dt = \frac{4(-1)^n}{n^2}. \end{aligned}$$

Note that we have “detected” the continuity of the extension since the coefficients decay as $\frac{1}{n^2}$. That is, the even periodic extension of t^2 has no jump discontinuities. It does have corners, since the derivative, which is an odd function and a sine series, has jumps; it has a Fourier series whose coefficients decay only as $\frac{1}{n}$.

Explicitly, the first few terms of the series are

$$\frac{\pi^2}{3} - 4 \cos(t) + \cos(2t) - \frac{4}{9} \cos(3t) + \dots$$

□

Exercise 9.4.3:

- a) Compute the derivative of the even periodic extension of $f(t)$ above and verify it has jump discontinuities. Use the actual definition of $f(t)$, not its cosine series!
- b) Why is it that the derivative of the even periodic extension of $f(t)$ is the odd periodic extension of $f'(t)$?

9.4.3 Application

Fourier series ties in to the boundary value problems we studied earlier. Let us see this connection in an application.

Consider the boundary value problem for $0 < t < L$,

$$x''(t) + \lambda x(t) = f(t),$$

for the *Dirichlet boundary conditions* $x(0) = 0$, $x(L) = 0$. The Fredholm alternative ([Theorem 9.1.2](#) on page 452) says that as long as λ is not an eigenvalue of the underlying homogeneous problem, there exists a unique solution. Eigenfunctions of this eigenvalue problem are the functions $\sin\left(\frac{n\pi}{L}t\right)$. Therefore, to find the solution, we first find the Fourier sine series for $f(t)$. We write x also as a sine series, but with unknown coefficients. We substitute the series for x into the equation and solve for the unknown coefficients. If we have the *Neumann boundary conditions* $x'(0) = 0$, $x'(L) = 0$, we do the same procedure using the cosine series.

Let us see how this method works on examples.

Example 9.4.3: Take the boundary value problem for $0 < t < 1$,

$$x''(t) + 2x(t) = f(t),$$

where $f(t) = t$ on $0 < t < 1$, and satisfying the Dirichlet boundary conditions $x(0) = 0$, $x(1) = 0$.

Solution: We write $f(t)$ as a sine series

$$f(t) = \sum_{n=1}^{\infty} c_n \sin(n\pi t).$$

Compute

$$c_n = 2 \int_0^1 t \sin(n\pi t) dt = \frac{2(-1)^{n+1}}{n\pi}.$$

We write $x(t)$ as

$$x(t) = \sum_{n=1}^{\infty} b_n \sin(n\pi t).$$

We plug in to obtain

$$\begin{aligned} x''(t) + 2x(t) &= \underbrace{\sum_{n=1}^{\infty} -b_n n^2 \pi^2 \sin(n\pi t)}_{x''} + 2 \underbrace{\sum_{n=1}^{\infty} b_n \sin(n\pi t)}_{x} \\ &= \sum_{n=1}^{\infty} b_n (2 - n^2 \pi^2) \sin(n\pi t) \\ &= f(t) = \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n\pi} \sin(n\pi t). \end{aligned}$$

Therefore,

$$b_n(2 - n^2\pi^2) = \frac{2(-1)^{n+1}}{n\pi}$$

or

$$b_n = \frac{2(-1)^{n+1}}{n\pi(2 - n^2\pi^2)}.$$

That $2 - n^2\pi^2$ is not zero for any n , and that we can solve for b_n , is precisely because 2 is not an eigenvalue of the problem. We have thus obtained a Fourier series for the solution

$$x(t) = \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n\pi(2 - n^2\pi^2)} \sin(n\pi t).$$

See Figure 9.13 for a graph of the solution. Notice that because the eigenfunctions satisfy the boundary conditions, and x is written in terms of the boundary conditions, then x satisfies the boundary conditions.

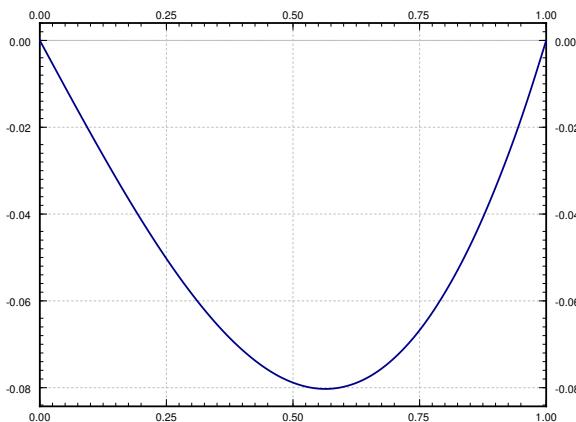


Figure 9.13: Plot of the solution of $x'' + 2x = t$, $x(0) = 0$, $x(1) = 0$.

Example 9.4.4: Similarly we handle the Neumann conditions. Take the boundary value problem for $0 < t < 1$,

$$x''(t) + 2x(t) = f(t),$$

where again $f(t) = t$ on $0 < t < 1$, but now satisfying the Neumann boundary conditions $x'(0) = 0$, $x'(1) = 0$.

Solution: We write $f(t)$ as a cosine series

$$f(t) = \frac{c_0}{2} + \sum_{n=1}^{\infty} c_n \cos(n\pi t),$$

where

$$c_0 = 2 \int_0^1 t dt = 1,$$

and

$$c_n = 2 \int_0^1 t \cos(n\pi t) dt = \frac{2((-1)^n - 1)}{\pi^2 n^2} = \begin{cases} \frac{-4}{\pi^2 n^2} & \text{if } n \text{ odd,} \\ 0 & \text{if } n \text{ even.} \end{cases}$$

We write $x(t)$ as a cosine series

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\pi t).$$

We plug in to obtain

$$\begin{aligned} x''(t) + 2x(t) &= \sum_{n=1}^{\infty} [-a_n n^2 \pi^2 \cos(n\pi t)] + a_0 + 2 \sum_{n=1}^{\infty} [a_n \cos(n\pi t)] \\ &= a_0 + \sum_{n=1}^{\infty} a_n (2 - n^2 \pi^2) \cos(n\pi t) \\ &= f(t) = \frac{1}{2} + \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{-4}{\pi^2 n^2} \cos(n\pi t). \end{aligned}$$

Therefore, $a_0 = \frac{1}{2}$, $a_n = 0$ for n even ($n \geq 2$) and for n odd we have

$$a_n (2 - n^2 \pi^2) = \frac{-4}{\pi^2 n^2},$$

or

$$a_n = \frac{-4}{n^2 \pi^2 (2 - n^2 \pi^2)}.$$

The Fourier series for the solution $x(t)$ is

$$x(t) = \frac{1}{4} + \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{-4}{n^2 \pi^2 (2 - n^2 \pi^2)} \cos(n\pi t).$$

□

9.4.4 Exercises

Exercise 9.4.4: Take $f(t) = (t - 1)^2$ defined on $0 \leq t \leq 1$.

a) Sketch the plot of the even periodic extension of f .

b) Sketch the plot of the odd periodic extension of f .

Exercise 9.4.5: Find the Fourier series of both the odd and even periodic extension of the function $f(t) = (t - 1)^2$ for $0 \leq t \leq 1$. Can you tell which extension is continuous from the Fourier series coefficients?

Exercise 9.4.6: Find the Fourier series of both the odd and even periodic extension of the function $f(t) = t$ for $0 \leq t \leq \pi$.

Exercise 9.4.7:* Let $f(t) = t/3$ on $0 \leq t < 3$.

- a) Find the Fourier series of the even periodic extension.
- b) Find the Fourier series of the odd periodic extension.

Exercise 9.4.8: Find the Fourier series of the even periodic extension of the function $f(t) = \sin t$ for $0 \leq t \leq \pi$.

Exercise 9.4.9:* Let $f(t) = \cos(2t)$ on $0 \leq t < \pi$.

- a) Find the Fourier series of the even periodic extension.
- b) Find the Fourier series of the odd periodic extension.

Exercise 9.4.10:* Let $f(t)$ be defined on $0 \leq t < 1$. Now take the average of the two extensions $g(t) = \frac{F_{\text{odd}}(t) + F_{\text{even}}(t)}{2}$.

- a) What is $g(t)$ if $0 \leq t < 1$ (Justify!)
- b) What is $g(t)$ if $-1 < t < 0$ (Justify!)

Exercise 9.4.11: Consider

$$x''(t) + 4x(t) = f(t),$$

where $f(t) = 1$ on $0 < t < 1$.

- a) Solve for the Dirichlet conditions $x(0) = 0, x(1) = 0$.
- b) Solve for the Neumann conditions $x'(0) = 0, x'(1) = 0$.

Exercise 9.4.12: Consider

$$x''(t) + 9x(t) = f(t),$$

for $f(t) = \sin(2\pi t)$ on $0 < t < 1$.

- a) Solve for the Dirichlet conditions $x(0) = 0, x(1) = 0$.
- b) Solve for the Neumann conditions $x'(0) = 0, x'(1) = 0$.

Exercise 9.4.13:* Let $f(t) = \sum_{n=1}^{\infty} \frac{1}{n^2} \sin(nt)$. Solve $x'' - x = f(t)$ for the Dirichlet conditions $x(0) = 0$ and $x(\pi) = 0$.

Exercise 9.4.14: Consider

$$x''(t) + 3x(t) = f(t), \quad x(0) = 0, \quad x(1) = 0,$$

where $f(t) = \sum_{n=1}^{\infty} b_n \sin(n\pi t)$. Write the solution $x(t)$ as a Fourier series, where the coefficients are given in terms of b_n .

Exercise 9.4.15: Let $f(t) = t^2(2-t)$ for $0 \leq t \leq 2$. Let $F(t)$ be the odd periodic extension. Compute $F(1), F(2), F(3), F(-1), F(\frac{9}{2}), F(101), F(103)$. Note: Do **not** compute using the sine series.

Exercise 9.4.16 (challenging):* Let $f(t) = t + \sum_{n=1}^{\infty} \frac{1}{2^n} \sin(nt)$. Solve $x'' + \pi x = f(t)$ for the Dirichlet conditions $x(0) = 0$ and $x(\pi) = 1$. Hint: Note that $\frac{t}{\pi}$ satisfies the given Dirichlet conditions.

9.5 Applications of Fourier series

Attribution: [JL], §4.5.

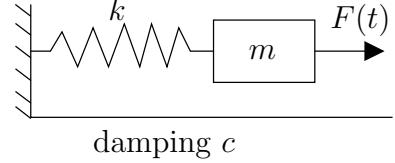
Learning Objectives

After this section, you will be able to:

- Apply Fourier Series to solve forced oscillation problems, and
- Understand how resonance shows up in these types of problems.

9.5.1 Periodically forced oscillation

Let us return to the forced oscillations. Consider a mass-spring system as before, where we have a mass m on a spring with spring constant k , with damping c , and a force $F(t)$ applied to the mass. Suppose the forcing function $F(t)$ is $2L$ -periodic for some $L > 0$. We saw this problem in chapter 2 with $F(t) = F_0 \cos(\omega t)$. The equation that governs this particular setup is



$$mx''(t) + cx'(t) + kx(t) = F(t). \quad (9.9)$$

The general solution of (9.9) consists of the complementary solution x_c , which solves the associated homogeneous equation $mx'' + cx' + kx = 0$, and a particular solution of (9.9) we call x_p . For $c > 0$, the complementary solution x_c will decay as time goes by. Therefore, we are mostly interested in a particular solution x_p that does not decay and is periodic with the same period as $F(t)$. We call this particular solution the *steady periodic solution* and we write it as x_{sp} as before. What is new in this section is that we consider an arbitrary forcing function $F(t)$ instead of a simple cosine.

For simplicity, suppose $c = 0$. The problem with $c > 0$ is very similar. The equation

$$mx'' + kx = 0$$

has the general solution

$$x(t) = A \cos(\omega_0 t) + B \sin(\omega_0 t),$$

where $\omega_0 = \sqrt{\frac{k}{m}}$. Any solution to $mx''(t) + kx(t) = F(t)$ is of the form $A \cos(\omega_0 t) + B \sin(\omega_0 t) + x_{sp}$. The steady periodic solution x_{sp} has the same period as $F(t)$.

In the spirit of the last section and the idea of undetermined coefficients we first write

$$F(t) = \frac{c_0}{2} + \sum_{n=1}^{\infty} c_n \cos\left(\frac{n\pi}{L}t\right) + d_n \sin\left(\frac{n\pi}{L}t\right).$$

Then we write a proposed steady periodic solution x as

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}t\right) + b_n \sin\left(\frac{n\pi}{L}t\right),$$

where a_n and b_n are unknowns. We plug x into the differential equation and solve for a_n and b_n in terms of c_n and d_n . This process is perhaps best understood by example.

Example 9.5.1: Suppose that $k = 2$, and $m = 1$. The units are again the mks units (meters-kilograms-seconds). There is a jetpack strapped to the mass, which fires with a force of 1 newton for 1 second and then is off for 1 second, and so on. We want to find the steady periodic solution.

Solution: The equation is, therefore,

$$x'' + 2x = F(t),$$

where $F(t)$ is the step function

$$F(t) = \begin{cases} 0 & \text{if } -1 < t < 0, \\ 1 & \text{if } 0 < t < 1, \end{cases}$$

extended periodically. We write

$$F(t) = \frac{c_0}{2} + \sum_{n=1}^{\infty} c_n \cos(n\pi t) + d_n \sin(n\pi t).$$

We compute

$$\begin{aligned} c_n &= \int_{-1}^1 F(t) \cos(n\pi t) dt = \int_0^1 \cos(n\pi t) dt = 0 \quad \text{for } n \geq 1, \\ c_0 &= \int_{-1}^1 F(t) dt = \int_0^1 dt = 1, \\ d_n &= \int_{-1}^1 F(t) \sin(n\pi t) dt \\ &= \int_0^1 \sin(n\pi t) dt \\ &= \left[\frac{-\cos(n\pi t)}{n\pi} \right]_{t=0}^1 \\ &= \frac{1 - (-1)^n}{\pi n} = \begin{cases} \frac{2}{\pi n} & \text{if } n \text{ odd,} \\ 0 & \text{if } n \text{ even.} \end{cases} \end{aligned}$$

So

$$F(t) = \frac{1}{2} + \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{2}{\pi n} \sin(n\pi t).$$

We want to try

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\pi t) + b_n \sin(n\pi t).$$

Once we plug x into the differential equation $x'' + 2x = F(t)$, it is clear that $a_n = 0$ for $n \geq 1$ as there are no corresponding terms in the series for $F(t)$. Similarly $b_n = 0$ for n even. Hence we try

$$x(t) = \frac{a_0}{2} + \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} b_n \sin(n\pi t).$$

We plug into the differential equation and obtain

$$\begin{aligned} x'' + 2x &= \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \left[-b_n n^2 \pi^2 \sin(n\pi t) \right] + a_0 + 2 \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \left[b_n \sin(n\pi t) \right] \\ &= a_0 + \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} b_n (2 - n^2 \pi^2) \sin(n\pi t) \\ &= F(t) = \frac{1}{2} + \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{2}{\pi n} \sin(n\pi t). \end{aligned}$$

So $a_0 = \frac{1}{2}$, $b_n = 0$ for even n , and for odd n we get

$$b_n = \frac{2}{\pi n (2 - n^2 \pi^2)}.$$

The steady periodic solution has the Fourier series

$$x_{sp}(t) = \frac{1}{4} + \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{2}{\pi n (2 - n^2 \pi^2)} \sin(n\pi t).$$

We know this is the steady periodic solution as it contains no terms of the complementary solution and it is periodic with the same period as $F(t)$ itself. See Figure 9.14 on the next page for the plot of this solution. \square

9.5.2 Resonance

Just as when the forcing function was a simple cosine, we may encounter resonance. Assume $c = 0$ and let us discuss only pure resonance. Let $F(t)$ be $2L$ -periodic and consider

$$mx''(t) + kx(t) = F(t).$$

When we expand $F(t)$ and find that some of its terms coincide with the complementary solution to $mx'' + kx = 0$, we cannot use those terms in the guess. Just like before, they disappear when we plug them into the left-hand side and we get a contradictory equation (such as $0 = 1$). That is, suppose

$$x_c = A \cos(\omega_0 t) + B \sin(\omega_0 t),$$

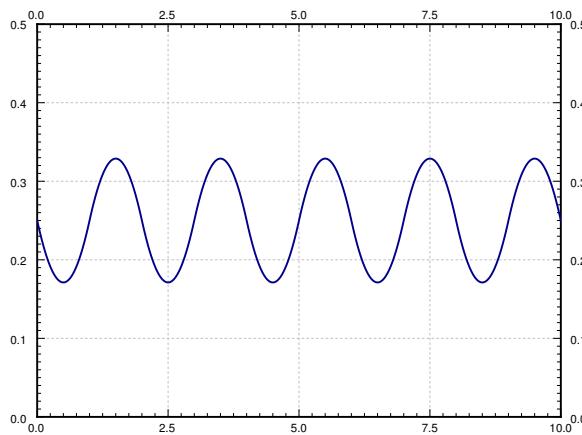


Figure 9.14: Plot of the steady periodic solution x_{sp} of Example 9.5.1.

where $\omega_0 = \frac{N\pi}{L}$ for some positive integer N . We have to modify our guess and try

$$x(t) = \frac{a_0}{2} + t \left(a_N \cos \left(\frac{N\pi}{L} t \right) + b_N \sin \left(\frac{N\pi}{L} t \right) \right) + \sum_{\substack{n=1 \\ n \neq N}}^{\infty} a_n \cos \left(\frac{n\pi}{L} t \right) + b_n \sin \left(\frac{n\pi}{L} t \right).$$

In other words, we multiply the offending term by t . From then on, we proceed as before.

Of course, the solution is not a Fourier series (it is not even periodic) since it contains these terms multiplied by t . Further, the terms $t(a_N \cos(\frac{N\pi}{L}t) + b_N \sin(\frac{N\pi}{L}t))$ eventually dominate and lead to wild oscillations. As before, this behavior is called *pure resonance* or just *resonance*.

Note that there now may be infinitely many resonance frequencies to hit. That is, as we change the frequency of F (we change L), different terms from the Fourier series of F may interfere with the complementary solution and cause resonance. However, we should note that since everything is an approximation and in particular c is never actually zero but something very close to zero, only the first few resonance frequencies matter in real life.

Example 9.5.2: We want to solve the equation

$$2x'' + 18\pi^2 x = F(t), \quad (9.10)$$

where

$$F(t) = \begin{cases} -1 & \text{if } -1 < t < 0, \\ 1 & \text{if } 0 < t < 1, \end{cases}$$

extended periodically.

Solution: We note that

$$F(t) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{4}{\pi n} \sin(n\pi t).$$

Exercise 9.5.1: Compute the Fourier series of F to verify the equation above.

As $\sqrt{\frac{k}{m}} = \sqrt{\frac{18\pi^2}{2}} = 3\pi$, the solution to (9.10) is

$$x(t) = c_1 \cos(3\pi t) + c_2 \sin(3\pi t) + x_p(t)$$

for some particular solution x_p .

If we just try an x_p given as a Fourier series with $\sin(n\pi t)$ as usual, the complementary equation, $2x'' + 18\pi^2 x = 0$, eats our 3rd harmonic. That is, the term with $\sin(3\pi t)$ is already in our complementary solution. Therefore, we pull that term out and multiply it by t . We also add a cosine term to get everything right. That is, we try

$$x_p(t) = a_3 t \cos(3\pi t) + b_3 t \sin(3\pi t) + \sum_{\substack{n=1 \\ n \text{ odd} \\ n \neq 3}}^{\infty} b_n \sin(n\pi t).$$

Let us compute the second derivative.

$$\begin{aligned} x_p''(t) &= -6a_3\pi \sin(3\pi t) - 9\pi^2 a_3 t \cos(3\pi t) + 6b_3\pi \cos(3\pi t) - 9\pi^2 b_3 t \sin(3\pi t) \\ &\quad + \sum_{\substack{n=1 \\ n \text{ odd} \\ n \neq 3}}^{\infty} (-n^2\pi^2 b_n) \sin(n\pi t). \end{aligned}$$

We now plug into the left-hand side of the differential equation.

$$\begin{aligned} 2x_p'' + 18\pi^2 x_p &= -12a_3\pi \sin(3\pi t) - 18\pi^2 a_3 t \cos(3\pi t) + 12b_3\pi \cos(3\pi t) - 18\pi^2 b_3 t \sin(3\pi t) \\ &\quad + 18\pi^2 a_3 t \cos(3\pi t) + 18\pi^2 b_3 t \sin(3\pi t) \\ &\quad + \sum_{\substack{n=1 \\ n \text{ odd} \\ n \neq 3}}^{\infty} (-2n^2\pi^2 b_n + 18\pi^2 b_n) \sin(n\pi t). \end{aligned}$$

We simplify,

$$2x_p'' + 18\pi^2 x_p = -12a_3\pi \sin(3\pi t) + 12b_3\pi \cos(3\pi t) + \sum_{\substack{n=1 \\ n \text{ odd} \\ n \neq 3}}^{\infty} (-2n^2\pi^2 b_n + 18\pi^2 b_n) \sin(n\pi t).$$

This series has to equal to the series for $F(t)$. We equate the coefficients and solve for a_3 and b_n .

$$a_3 = \frac{4/(3\pi)}{-12\pi} = \frac{-1}{9\pi^2},$$

$$b_3 = 0,$$

$$b_n = \frac{4}{n\pi(18\pi^2 - 2n^2\pi^2)} = \frac{2}{\pi^3 n(9 - n^2)} \quad \text{for } n \text{ odd and } n \neq 3.$$

That is,

$$x_p(t) = \frac{-1}{9\pi^2} t \cos(3\pi t) + \sum_{\substack{n=1 \\ n \text{ odd} \\ n \neq 3}}^{\infty} \frac{2}{\pi^3 n(9 - n^2)} \sin(n\pi t).$$

□

When $c > 0$, you do not have to worry about pure resonance. That is, there are never any conflicts and you do not need to multiply any terms by t . There is a corresponding concept of practical resonance and it is very similar to the ideas we already explored in [chapter 2](#). Basically what happens in practical resonance is that one of the coefficients in the series for x_{sp} can get very big. Let us not go into details here.

9.5.3 Exercises

Exercise 9.5.2: Let $F(t) = \frac{1}{2} + \sum_{n=1}^{\infty} \frac{1}{n^2} \cos(n\pi t)$. Find the steady periodic solution to $x'' + 2x = F(t)$. Express your solution as a Fourier series.

Exercise 9.5.3:* Let $F(t) = \sin(2\pi t) + 0.1 \cos(10\pi t)$. Find the steady periodic solution to $x'' + \sqrt{2}x = F(t)$. Express your solution as a Fourier series.

Exercise 9.5.4: Let $F(t) = \sum_{n=1}^{\infty} \frac{1}{n^3} \sin(n\pi t)$. Find the steady periodic solution to $x'' + x' + x = F(t)$. Express your solution as a Fourier series.

Exercise 9.5.5:* Let $F(t) = \sum_{n=1}^{\infty} e^{-n} \cos(2nt)$. Find the steady periodic solution to $x'' + 3x = F(t)$. Express your solution as a Fourier series.

Exercise 9.5.6: Let $F(t) = \sum_{n=1}^{\infty} \frac{1}{n^2} \cos(n\pi t)$. Find the steady periodic solution to $x'' + 4x = F(t)$. Express your solution as a Fourier series.

Exercise 9.5.7: Let $F(t) = t$ for $-1 < t < 1$ and extended periodically. Find the steady periodic solution to $x'' + x = F(t)$. Express your solution as a series.

Exercise 9.5.8:* Let $F(t) = |t|$ for $-1 \leq t \leq 1$ extended periodically. Find the steady periodic solution to $x'' + \sqrt{3}x = F(t)$. Express your solution as a series.

Exercise 9.5.9: Let $F(t) = t$ for $-1 < t < 1$ and extended periodically. Find the steady periodic solution to $x'' + \pi^2 x = F(t)$. Express your solution as a series.

Exercise 9.5.10:* Let $F(t) = |t|$ for $-1 \leq t \leq 1$ extended periodically. Find the steady periodic solution to $x'' + \pi^2 x = F(t)$. Express your solution as a series.

Chapter 10

Introduction to PDEs

10.1 First order linear PDE

Attribution: [JL], §1.9.

Learning Objectives

After this section, you will be able to:

- Identify first order linear PDE and
- Use the method of characteristics to solve first order PDE.

We begin this chapter with an introduction to PDE in general, before moving on to techniques involving the Fourier Series discussed in [Chapter 9](#).

Consider the equation

$$a(x, t) u_x + b(x, t) u_t + c(x, t) u = g(x, t), \quad u(x, 0) = f(x), \quad -\infty < x < \infty, \quad t > 0,$$

where $u(x, t)$ is a function of x and t . The *initial condition* $u(x, 0) = f(x)$ is now a function of x rather than just a number. In these problems, it is useful to think of x as position and t as time. The equation describes the evolution of a function of x as time goes on. Below, the coefficients a , b , c , and the function g are mostly going to be constant or zero. The method we describe works with nonconstant coefficients, although the computations may get difficult quickly.

This method we use is the *method of characteristics*. The idea is that we find lines along which the equation is an ODE that we solve. We will see this technique again for second order PDE when we encounter the wave equation in [§ 10.5](#).

Example 10.1.1: Consider the equation

$$u_t + \alpha u_x = 0, \quad u(x, 0) = f(x).$$

This particular equation, $u_t + \alpha u_x = 0$, is called the *transport equation*.

Solution: The data will propagate along curves called characteristics. The idea is to change to the so-called *characteristic coordinates*. If we change to these coordinates, the equation

simplifies. The change of variables for this equation is

$$\xi = x - \alpha t, \quad s = t.$$

Let's see what the equation becomes. Remember the chain rule in several variables.

$$\begin{aligned} u_t &= u_\xi \xi_t + u_s s_t = -\alpha u_\xi + u_s, \\ u_x &= u_\xi \xi_x + u_s s_x = u_\xi. \end{aligned}$$

The equation in the coordinates ξ and s becomes

$$\underbrace{(-\alpha u_\xi + u_s)}_{u_t} + \alpha \underbrace{(u_\xi)}_{u_x} = 0,$$

or in other words

$$u_s = 0.$$

That is trivial to solve. Treating ξ as simply a parameter, we have obtained the ODE $\frac{du}{ds} = 0$.

The solution is a function that does not depend on s (but it does depend on ξ). That is, there is some function A such that

$$u = A(\xi) = A(x - \alpha t).$$

The initial condition says that:

$$f(x) = u(x, 0) = A(x - \alpha 0) = A(x),$$

so $A = f$. In other words,

$$u(x, t) = f(x - \alpha t).$$

Everything is simply moving right at speed α as t increases. The curve given by the equation

$$\xi = \text{constant}$$

is called the characteristic. See [Figure 10.1](#). In this case, the solution does not change along the characteristic.

In the (x, t) coordinates, the characteristic curves satisfy $t = \frac{1}{\alpha}(x - \xi)$, and are in fact lines. The slope of characteristic lines is $\frac{1}{\alpha}$, and for each different ξ we get a different characteristic line.

We see why $u_t + \alpha u_x = 0$ is called the transport equation: everything travels at some constant speed. Sometimes this is called *convection*. An example application is material being moved by a river where the material does not diffuse and is simply carried along. In this setup, x is the position along the river, t is the time, and $u(x, t)$ the concentration the material at position x and time t . See [Figure 10.2](#) on the next page for an example. □

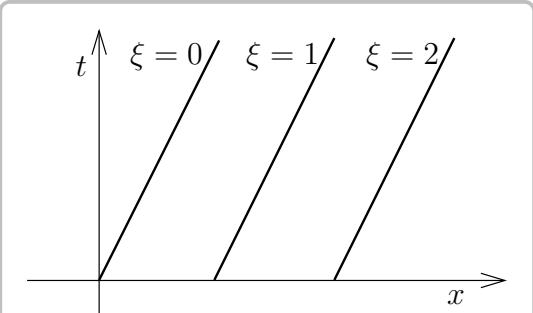


Figure 10.1: Characteristic curves.

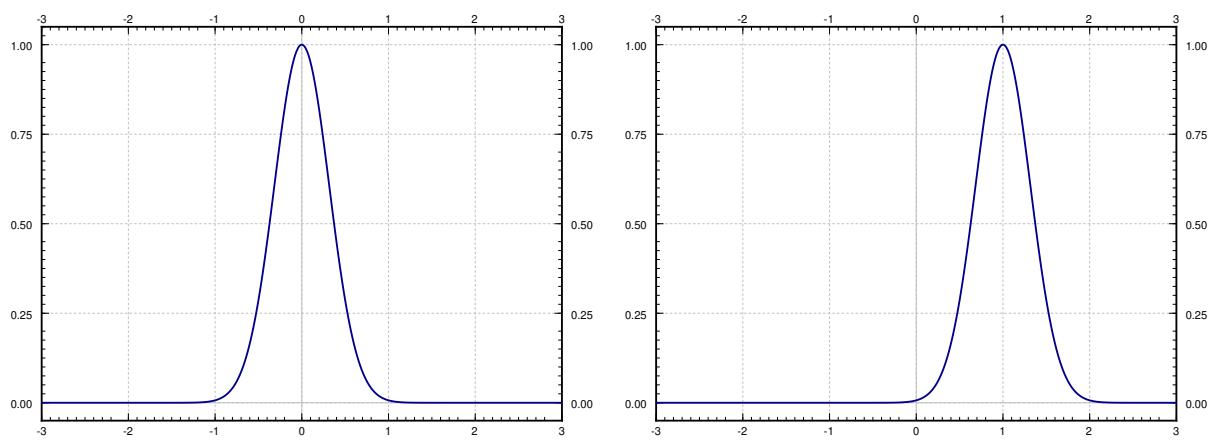


Figure 10.2: Example of “transport” in $u_t - u_x = 0$ (that is, $\alpha = 1$) where the initial condition $f(x)$ is a peak at the origin. On the left is a graph of the initial condition $u(x, 0)$. On the right is a graph of the function $u(x, 1)$, that is at time $t = 1$. Notice it is the same graph shifted one unit to the right.

We use similar idea in the more general case:

$$au_x + bu_t + cu = g, \quad u(x, 0) = f(x).$$

We change coordinates to the characteristic coordinates. Let us call these coordinates (ξ, s) . These are coordinates where $au_x + bu_t$ becomes differentiation in the s variable.

Along the characteristic curves (where ξ is constant), we get a new ODE in the s variable. In the transport equation, we got the simple $\frac{du}{ds} = 0$. In general, we get the linear equation

$$\frac{du}{ds} + cu = g. \quad (10.1)$$

We think of everything as a function of ξ and s , although we are thinking of ξ as a parameter rather than an independent variable. So the equation is an ODE. It is a linear ODE that we can solve using the integrating factor.

To find the characteristics, think of a curve given parametrically $(x(s), t(s))$. We try to have the curve satisfy

$$\frac{dx}{ds} = a, \quad \frac{dt}{ds} = b.$$

Why? Because when we think of x and t as functions of s we find, using the chain rule,

$$\frac{du}{ds} + cu = \underbrace{\left(u_x \frac{dx}{ds} + u_t \frac{dt}{ds} \right)}_{\frac{du}{ds}} + cu = au_x + bu_t + cu = g.$$

So we get the ODE (10.1), which then describes the value of the solution u of the PDE along this characteristic curve. It is also convenient to make sure that $s = 0$ corresponds to $t = 0$,

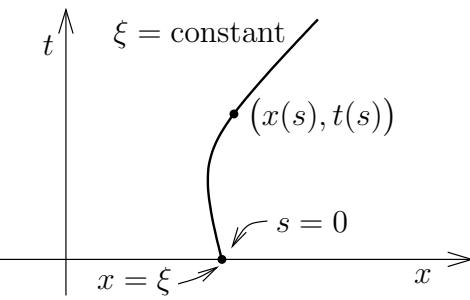


Figure 10.3: General characteristic curve.

that is $t(0) = 0$. It will be convenient also for $x(0) = \xi$. See [Figure 10.3](#) on the following page.

Example 10.1.2: Consider

$$u_x + u_t + u = x, \quad u(x, 0) = e^{-x^2}.$$

Solution: We find the characteristics, that is, the curves given by

$$\frac{dx}{ds} = 1, \quad \frac{dt}{ds} = 1.$$

So

$$x = s + c_1, \quad t = s + c_2,$$

for some c_1 and c_2 . At $s = 0$ we want $t = 0$, and x should be ξ . So we let $c_1 = \xi$ and $c_2 = 0$:

$$x = s + \xi, \quad t = s.$$

The ODE is $\frac{du}{ds} + u = x$, and $x = s + \xi$. So, the ODE to solve along the characteristic is

$$\frac{du}{ds} + u = s + \xi.$$

The general solution of this equation, treating ξ as a parameter, is $u = Ce^{-s} + s + \xi - 1$, for some constant C . At $s = 0$, our initial condition is that u is $e^{-\xi^2}$, since at $s = 0$ we have $x = \xi$. Given this initial condition, we find $C = e^{-\xi^2} - \xi + 1$. So,

$$\begin{aligned} u &= (e^{-\xi^2} - \xi + 1)e^{-s} + s + \xi - 1 \\ &= e^{-\xi^2-s} + (1 - \xi)e^{-s} + s + \xi - 1. \end{aligned}$$

Substitute $\xi = x - t$ and $s = t$ to find u in terms of x and t :

$$\begin{aligned} u &= e^{-\xi^2-s} + (1 - \xi)e^{-s} + s + \xi - 1 \\ &= e^{-(x-t)^2-t} + (1 - x + t)e^{-t} + x - 1. \end{aligned}$$

See [Figure 10.4](#) on the next page for a plot of $u(x, t)$ as a function of two variables.

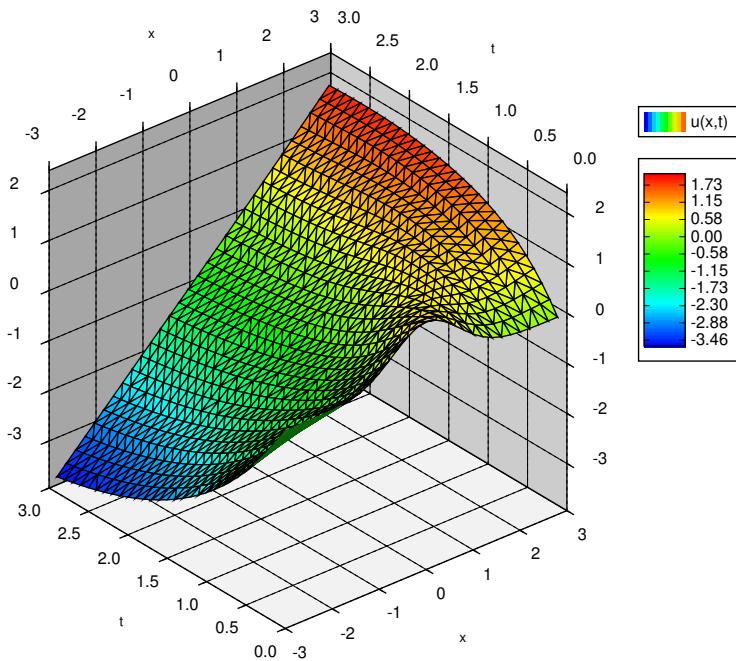


Figure 10.4: Plot of the solution $u(x,t)$ to $u_x + u_t + u = x$, $u(x,0) = e^{-x^2}$.

When the coefficients are not constants, the characteristic curves are not going to be straight lines anymore.

Example 10.1.3: Consider the following variable coefficient equation:

$$xu_x + u_t + 2u = 0, \quad u(x,0) = \cos(x).$$

Solution: We find the characteristics, that is, the curves given by

$$\frac{dx}{ds} = x, \quad \frac{dt}{ds} = 1.$$

So

$$x = c_1 e^s, \quad t = s + c_2.$$

At $s = 0$, we wish to get the line $t = 0$, and x should be ξ . So

$$x = \xi e^s, \quad t = s.$$

OK, the ODE we need to solve is

$$\frac{du}{ds} + 2u = 0.$$

This is for a fixed ξ . At $s = 0$, we should get that u is $\cos(\xi)$, so that is our initial condition. Consequently,

$$u = e^{-2s} \cos(\xi) = e^{-2t} \cos(xe^{-t}).$$

_

We make a few closing remarks. One thing to keep in mind is that we would get into trouble if the coefficient in front of u_t , that is the b , is ever zero. Let us consider a quick example of what can go wrong:

$$u_x + u = 0, \quad u(x, 0) = \sin(x).$$

This problem has no solution. If we had a solution, it would imply that $u_x(x, 0) = \cos(x)$, but $u_x(x, 0) + u(x, 0) = \cos(x) + \sin(x) \neq 0$. The problem is that the characteristic curve is now the line $t = 0$, and the solution is already provided on that line!

As long as b is nonzero, it is convenient to ensure that b is positive by multiplying by -1 if necessary, so that positive s means positive t .

Another remark is that if a or b in the equation are variable, the computations can quickly get out of hand, as the expressions for the characteristic coordinates become messy and then solving the ODE becomes even messier. In the examples above, b was always 1, meaning we got $s = t$ in the characteristic coordinates. If b is not constant, your expression for s will be more complicated.

Finding the characteristic coordinates is really a system of ODE in general if a depends on t or if b depends on x . In that case, we would need techniques of systems of ODE to solve, see [chapter 4](#) or [chapter 5](#). In general, if a and b are not linear functions or constants, finding closed form expressions for the characteristic coordinates may be impossible.

Finally, the method of characteristics applies to nonlinear first order PDE as well. In the nonlinear case, the characteristics depend not only on the differential equation, but also on the initial data. This leads to not only more difficult computations, but also the formation of singularities where the solution breaks down at a certain point in time. An example application where first order nonlinear PDE come up is traffic flow theory, and you have probably experienced the formation of singularities: traffic jams. But we digress.

10.1.1 Exercises

Exercise 10.1.1: Solve

- | | |
|--|--|
| a) $u_t + 9u_x = 0, \quad u(x, 0) = \sin(x),$ | b) $u_t - 8u_x = 0, \quad u(x, 0) = \sin(x),$ |
| c) $u_t + \pi u_x = 0, \quad u(x, 0) = \sin(x),$ | d) $u_t + \pi u_x + u = 0, \quad u(x, 0) = \sin(x).$ |

Exercise 10.1.2:* Solve

- | | |
|---|---|
| a) $u_t - 5u_x = 0, \quad u(x, 0) = \frac{1}{1+x^2},$ | b) $u_t + 2u_x = 0, \quad u(x, 0) = \cos(x).$ |
|---|---|

Exercise 10.1.3: Solve $u_t + 3u_x = 1, \quad u(x, 0) = x^2$.

Exercise 10.1.4:* Solve $u_x + u_t + tu = 0, \quad u(x, 0) = \cos(x)$.

Exercise 10.1.5: Solve $u_t + 3u_x = x, \quad u(x, 0) = e^x$.

Exercise 10.1.6: Solve $u_x + u_t + xu = 0$, $u(x, 0) = \cos(x)$.

Exercise 10.1.7:* Solve $u_x + u_t = 5$, $u(x, 0) = x$.

Exercise 10.1.8:

a) Find the characteristic coordinates for the following equations:

$$1) \quad u_x + u_t + u = 1, \quad u(x, 0) = \cos(x), \quad 2) \quad 2u_x + 2u_t + 2u = 2, \quad u(x, 0) = \cos(x).$$

b) Solve the two equations using the coordinates.

c) Explain why you got the same solution, although the characteristic coordinates you found were different.

Exercise 10.1.9: Solve $(1 + x^2)u_t + x^2u_x + e^xu = 0$, $u(x, 0) = 0$. Hint: Think a little out of the box.

10.2 Second order linear PDEs

Learning Objectives

After this section, you will be able to:

- Classify second order linear partial differential equations,
- Use separation of variables to attempt to find a solution to a partial differential equation

10.2.1 Classification

As with ordinary differential equations, second order partial differential equations have a lot of varied physical applications. The next few sections of the book will go into these applications, but we will first set up the appropriate terminology around these equations. Assume that we have a function u of two variables x and y . A PDE is said to be *linear* if the dependent variable and its derivatives appear at most to the first power and in no functions. The general second order linear PDE for u looks like

$$A(x, y) \frac{\partial^2 u}{\partial x^2} + B(x, y) \frac{\partial^2 u}{\partial x \partial y} + C(x, y) \frac{\partial^2 u}{\partial y^2} + D(x, y) \frac{\partial u}{\partial x} + E(x, y) \frac{\partial u}{\partial y} + F(x, y)u = G(x, y).$$

This expression contains all possible derivatives of u up to second order, multiplied by functions of the independent variables x and y . If we had a different number of variables, then we would get a different set of terms here.

Exercise 10.2.1: Write out the general second order linear ODE for a function $u(x, y, z)$ of three independent variables. How many terms are there?

It is also possible to consider second order non-linear PDE, but if ODE work taught us anything, it's that non-linear equations are *much* harder to solve than linear ones. As we will see going forward, we can only solve these linear ones in very specific cases, so there isn't too much worth in trying to extend to non-linear equations at this point.

In addition to the equation, we need to factor in the *initial conditions* and *boundary conditions* for the problem. Boundary conditions specify the value of the solution or its derivatives along the boundary of the region (in space), and initial conditions give such values at some initial time. A general way to identify these is that sometimes one of the variables is indicated as time, which is generally labeled as t and the problem only gives information about what is happening at the initial time value, because we want to see how the equation develops. This gives an initial condition. In the other case, the independent variables generally indicate spatial dimensions and information is known at both ends of the desired domain, and these give rise to boundary conditions. Sometimes such conditions are mixed together and we will refer to them simply as *side conditions*.

Example 10.2.1: An example of the heat equation in one dimension is

$$\frac{\partial u}{\partial t} = 4 \frac{\partial^2 u}{\partial x^2}$$

for the unknown function $u(t, x)$, with conditions

$$u(0, x) = x(1 - x)$$

$$u(t, 0) = 0 \quad u(t, 1) = 0.$$

The notation of the variables indicates that t is time and x is the position. The conditions given also indicate this, because we are given information about what happens at *both* endpoints (0 and 1) in the x direction, so these are boundary conditions, but only information about the starting value of $t = 0$ in terms of t , so this is an initial condition.

Even these general equations are very hard to solve, so just like in § 2.1, we will restrict to constant coefficient equations. Returning to our example of a function u of two variables x and y , the general constant coefficient second order linear PDE is

$$A \frac{\partial^2 u}{\partial x^2} + B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} + D \frac{\partial u}{\partial x} + E \frac{\partial u}{\partial y} + Fu = G(x, y). \quad (10.2)$$

This equation is said to be *homogeneous* if $G(x, y) = 0$ and *non-homogeneous* otherwise. It turns out that all equations of this type can be divided into three main groups based on the coefficients.

Definition 10.2.1

An equation of the form (10.2) is said to be

1. *hyperbolic* if $B^2 - 4AC > 0$
2. *parabolic* if $B^2 - 4AC = 0$
3. *elliptic* if $B^2 - 4AC < 0$.

Remark 10.2.1: You may recognize these terms from conic sections and this is not a coincidence. If you consider the equation

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 1,$$

the graph of this is a hyperbola if $B^2 - 4AC > 0$, a parabola if $B^2 - 4AC = 0$, and an ellipse (or circle) if $B^2 - 4AC < 0$.

It is well beyond the scope of this book, but PDE of the same “type” behave similarly. That is, there are certain characteristics of the solutions to these PDE that are true for all equations of the same type. This means that to get a general idea of how solutions to these equations behave, we only need to look at some examples of them. The next sections will do exactly that, describing the heat equation (parabolic), wave equation (hyperbolic), and Laplace’s equation (elliptic). These are very simple examples of these types of equation that we will be able to deal with, and they will show the general behavior of each of these types of equations.

Exercise 10.2.2: Verify that the equations below have the type described in the text above.

$$\frac{\partial u}{\partial t} = 4 \frac{\partial^2 u}{\partial x^2} \quad (\text{Heat equation})$$

$$\frac{\partial^2 u}{\partial t^2} = 9 \frac{\partial^2 u}{\partial x^2} \quad (\text{Wave equation})$$

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad (\text{Laplace's equation})$$

Note that in the first two examples, the variables are t and x , while in the last one, the variables are x and y .

10.2.2 Separation of variables

In general, just like for more complicated ODE, solving a generic PDE for an explicit solution is nearly impossible. However, we do have existence and uniqueness theorems for these kinds of equations. These theorems basically say that as long as the equation is “nice” and the boundary and initial conditions are “nice” enough, then the solution will exist, at least for a short time. The “nice” part here is important for these PDE. For ODE, all of our initial conditions or boundary conditions were single numbers, and no number (besides maybe zero) is nicer than any other. However, for PDE, our boundary and initial conditions are all functions, which can range greatly in how “nice” they are. For example, the function

$$f_1(x) = \begin{cases} 1 & -1 \leq x \leq 1 \\ 0 & |x| > 1 \end{cases}$$

is not as nice as

$$f_2(x) = |x|,$$

because the first is not continuous, and $f_3(x) = x^2$ is even nicer, because it is differentiable.

This idea of “niceness” is usually referred to as the *regularity* of the solution, which refers to how many continuous derivatives a function has. Some of these differential equations (elliptic and parabolic) have “smoothing” properties, which means that the solution is more regular than the boundary or initial conditions. We’ll get to these types of properties in their respective sections. The main fact that comes out of these theorems (that we will not discuss in detail) is that the solution is at least as regular as the boundary and initial conditions.

All of this goes to say that in hunting for a solution to these PDE, we just need to find a solution that works, and then the existence and uniqueness theorems will guarantee that this is the only solution. The argument here is the same as back in § 2.1; once we found a solution of a particular form, we know that this is *the* solution because of the theorems. Previously, we used solutions of the form $y(x) = e^{rx}$ to find solutions to second order ODE. The technique we want to use for PDE is called *separation of variables*.

Let’s use the heat equation as an example here. We have a function $u(t, x)$ that we want to have solve the PDE

$$\frac{\partial u}{\partial t} = 4 \frac{\partial^2 u}{\partial x^2}.$$

Assume that we have the boundary conditions $u(t, 0) = 0$, $u(t, 1) = 0$ and initial condition $u(0, x) = f(x)$ on the interval $0 \leq x \leq 1$. The main trick with this technique is to separate the variables; that is, we will assume that the solution $u(t, x)$ can be written as

$$u(t, x) = T(t)X(x)$$

where T is a function only of the time variable t , and X is a function only of the position x . In this way, the variables have been separated to live in different functions. Then we can plug this into the differential equation to see what T and X must satisfy. Since u is written as a product of functions of different variables, we have that

$$\frac{\partial u}{\partial t} = T'(t)X(x)$$

and

$$\frac{\partial^2 u}{\partial x^2} = T(t)X''(x).$$

Thus, the PDE gives that

$$T'(t)X(x) = 4T(t)X''(x).$$

We can rewrite this equation as

$$\frac{T'(t)}{4T(t)} = \frac{X''(x)}{X(x)}.$$

This expression here looks fairly complicated. However, the left hand side is only a function of t , and the right hand side is only a function of x . Furthermore, we want this to hold for *all* inputs x and t . Therefore, both sides of this expression must equal the same constant. If the left hand side changes at all with t , then we can just change the t value to adjust it and break the equality. Therefore, they are both constant, which we will call $-\lambda$, with the minus sign for convenience.

With this, we can separate the differential equation into two separate ODE for T and X respectively, giving

$$\frac{T'(t)}{4T(t)} = -\lambda \quad \frac{X''(x)}{X(x)} = -\lambda$$

which can be rewritten as

$$T'(t) + 4\lambda T(t) = 0 \quad X''(x) + \lambda X(x) = 0.$$

These are, respectively, a first and second order constant coefficient linear equation that can be solved by normal ODE methods.

The equation for $X(x)$ is very reminiscent of § 9.1, which is the eigenvalue-eigenfunction boundary value problem. At this point, λ is some constant, but we haven't specified the value yet. This value (or these possible values) will come as the eigenvalues of the problem with the appropriate boundary conditions. But what are these boundary conditions?

The boundary conditions for the PDE as a whole say that $u(t, 0) = 0$ and $u(t, 1) = 0$. If we write these out in terms of the separated functions, we get that

$$T(t)X(0) = 0 \quad T(t)X(1) = 0$$

for all values of t . Since we don't want $T(t) = 0$ (since that would make the entire function $u(t, x) = 0$), we must have that $X(0) = 0$ and $X(1) = 0$. This gives the eigenvalue problem

$$X'' + \lambda X = 0 \quad X(0) = 0, \quad X(1) = 0.$$

This problem can be solved using the methods of § 9.1, giving that we have eigenvalues of $\lambda_n = n^2\pi^2$ with eigenfunctions

$$X_n(x) = A_n \sin(n\pi x).$$

We can then carry these eigenvalues over to the $T(t)$ equation, needing to solve

$$T' + 4n^2\pi^2 T = 0$$

which has solution

$$T_n(t) = C_n e^{-4n^2\pi^2 t}.$$

Finally, we can combine these together to get

$$u_n(t, x) = A_n e^{-4n^2\pi^2 t} \sin(n\pi x),$$

all of which solve the original PDE and boundary conditions $u_n(t, 0) = 0$ and $u_n(t, 1) = 0$.

Exercise 10.2.3: Check that u_n meets all of these conditions for any positive integer n .

So that takes care of the PDE and the boundary conditions. But now we have a *bunch* of solutions, one for each n , and we have yet to meet the initial condition. Thankfully, the heat equation is linear as u and its derivatives do not appear to any powers or in any functions. Thus the principle of superposition still applies for the heat equation (without side conditions): If u_1 and u_2 are solutions and c_1, c_2 are constants, then $u = c_1 u_1 + c_2 u_2$ is also a solution.

Exercise 10.2.4: Verify the principle of superposition for the heat equation.

Superposition preserves some of the side conditions. In particular, if u_1 and u_2 are solutions that satisfy $u(0, t) = 0$ and $u(L, t) = 0$, and c_1, c_2 are constants, then $u = c_1 u_1 + c_2 u_2$ is still a solution that satisfies $u(0, t) = 0$ and $u(L, t) = 0$. Similarly for the side conditions $u_x(0, t) = 0$ and $u_x(L, t) = 0$. In general, superposition preserves all homogeneous side conditions.

This means that we can add up our various u_n to get additional solutions. In particular, for choices of constants A_n , the function

$$u(t, x) = \sum_{n=1}^{\infty} A_n u_n = \sum_{n=1}^{\infty} A_n e^{-4n^2\pi^2 t} \sin(n\pi x)$$

will solve the PDE and meet the boundary conditions. For this function, what happens at $t = 0$? If we set t to be zero, all of the exponential terms go to 1, leaving us with

$$u(0, x) = \sum_{n=1}^{\infty} A_n \sin(n\pi x)$$

which we want to match the function $f(x)$. The expression we had for $u(0, x)$ looks a lot like a Fourier series, in particular, a sine series. Thus, if we can expand the function f into a sine series of the form

$$f(x) = \sum_{n=1}^{\infty} A_n \sin(n\pi x)$$

this will give us the values of A_n that we should choose to construct our function $u(t, x)$. With those decisions made, we then have a function $u(t, x)$ that solves the PDE, the boundary conditions, and the initial condition. This means we have found the solution that we want!

This outlines the general process of using separation of variables to solve a PDE. The steps to this process are

1. Express the solution u as a product of functions of each of the different variables.
2. Plug this into the original differential equation and rewrite the equation so different variables are on different sides of the equation, so that each side of the equation must be a constant.
3. Separate the equations into different ODE for each of the independent variables in the equation, connected by the eigenvalue constants.
4. Translate the boundary conditions from the PDE into boundary conditions on the separated functions.
5. Solve the eigenvalue problem for the spatial dimension(s) to get the appropriate values of the constant.
6. Solve the rest of the ODE using these constants
7. Write a series to represent the solution to the full PDE using these separated solutions.
8. Use Fourier series to identify the constants needed and find the final solution to the PDE.

The following example will illustrate this process.

Example 10.2.2: Solve the PDE for the function $u(x, y)$ that solves

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} + 9 \frac{\partial^2 u}{\partial y^2} &= 0 \\ u(0, y) &= 0 \\ u(\pi, y) &= 0 \\ u(x, 0) &= \sin(x) \\ u(x, \pi) &= 0 \end{aligned}$$

Solution: In this case, we have an elliptic equation ($B^2 - 4AC = 0 - 8 < 0$), which is going to change what the solution looks like, but it won't substantially change the method we use to solve it. We start by looking for this building-block solutions in the form

$$u(x, y) = X(x)Y(y).$$

If we plug this into the differential equation, we get that these functions must solve

$$X''(x)Y(y) + 9X(x)Y''(y) = 0$$

or that

$$\frac{X''(x)}{9X(x)} = -\frac{Y''(y)}{Y(y)}.$$

Since these are functions of different variables, they must both equal the constant $-\lambda$. Thus, we get the two separated ODE

$$\frac{X''(x)}{9X(x)} = -\lambda \Rightarrow X''(x) + 9\lambda X(x) = 0$$

and

$$-\frac{Y''(y)}{Y(y)} = -\lambda \Rightarrow Y''(y) - \lambda Y(y) = 0.$$

Next, we want to consider the boundary conditions. Since this is an elliptic PDE, everything is a boundary condition, we could use any of them. The main components here that are helpful are the zero boundary conditions. Those tell us that

$$\begin{aligned} u(0, y) = 0 &\Rightarrow X(0) = 0 \\ u(\pi, y) = 0 &\Rightarrow X(\pi) = 0. \\ u(x, \pi) = 0 &\Rightarrow Y(\pi) = 0 \end{aligned}$$

Since we have two zero boundary conditions for the X equation, that boundary value problem is well-posed, and so it's something that we know how to solve. So, we start there. This boundary value problem is

$$X''(x) + 9\lambda X(x) = 0 \quad X(0) = 0, \quad X(\pi) = 0.$$

The solution of this eigenvalue problem is that $9\lambda = n^2$ for any positive integer n , with eigenfunction

$$X_n(x) = A_n \sin(nx).$$

We can then take the eigenvalue $\lambda = \frac{n^2}{9}$ and carry this over to the Y equation to get the equation

$$Y''(y) - \frac{n^2}{9}Y(y) = 0$$

which has general solution

$$Y_n(y) = C_1 e^{\frac{n}{3}y} + C_2 e^{-\frac{n}{3}y}.$$

With our separate functions solved, we can now stack them up into a series solution to the entire PDE. This gives that we are looking for the solution $u(x, y)$ in the form

$$u(x, y) = \sum_{n=1}^{\infty} A_n e^{\frac{n}{3}y} \sin(nx) + B_n e^{-\frac{n}{3}y} \sin(nx).$$

This function, as we have generated it, will solve the PDE, as well as the boundary conditions at $x = 0$ and $x = \pi$. To finish this, we need to pick the constants A_n and B_n so that we satisfy

$$u(x, 0) = \sin(x) \quad \text{and} \quad u(x, \pi) = 0.$$

Plugging $y = 0$ into the function, we get

$$u(x, 0) = \sum_{n=1}^{\infty} A_n \sin(nx) + B_n \sin(nx) = \sum_{n=1}^{\infty} (A_n + B_n) \sin(nx).$$

Since we want this to match the boundary condition, we would need to use Fourier series. However, this particular series is easy, because the boundary condition is already a sine function. Therefore we know that we need

$$A_1 + B_1 = 1 \quad A_n + B_n = 0 \quad n \geq 2$$

to give us just a $\sin(x)$ from this series. Next, we can plug $y = \pi$ into this series to get

$$u(x, \pi) = \sum_{n=1}^{\infty} A_n e^{\frac{n}{3}\pi} \sin(nx) + B_n e^{-\frac{n}{3}\pi} \sin(nx) = \sum_{n=1}^{\infty} (A_n e^{\frac{n}{3}\pi} + B_n e^{-\frac{n}{3}\pi}) \sin(nx).$$

We want this series to evaluate to zero, and since the functions $\sin(nx)$ are linearly independent, we need that

$$A_n e^{\frac{n}{3}\pi} + B_n e^{-\frac{n}{3}\pi} = 0$$

for all n . Since $e^{\frac{n}{3}\pi}$ and $e^{-\frac{n}{3}\pi}$ are different for all n , we get that for $n \geq 2$ the pair of equations

$$A_n + B_n = 0 \quad A_n e^{\frac{n}{3}\pi} + B_n e^{-\frac{n}{3}\pi} = 0$$

only has the solution $A_n = 0$ and $B_n = 0$. Therefore, the only interesting case here is $n = 1$, which gives rise to the pair of equations

$$A_1 + B_1 = 1 \quad A_1 e^{\frac{\pi}{3}} + B_1 e^{-\frac{\pi}{3}} = 0.$$

From the first equation, we can write that $A_1 = 1 - B_1$ and plug this into the second equation to get that

$$(1 - B_1) e^{\frac{\pi}{3}} + B_1 e^{-\frac{\pi}{3}} = 0,$$

which we can then solve for B_1 as

$$-B_1 e^{\frac{\pi}{3}} + B_1 e^{-\frac{\pi}{3}} = e^{\frac{\pi}{3}}$$

or

$$B_1 = \frac{e^{\frac{\pi}{3}}}{e^{-\frac{\pi}{3}} - e^{\frac{\pi}{3}}} = \frac{1}{e^{-\frac{2\pi}{3}} - 1}.$$

Then, we can solve for A_1 as

$$A_1 = 1 - B_1 = 1 - \frac{1}{e^{-\frac{2\pi}{3}} - 1}.$$

Finally, we plug these values for A_1 and B_1 , along with the fact that all of the other coefficients are zero, into the full solution to get

$$\begin{aligned} u(x, y) &= \sum_{n=1}^{\infty} A_n e^{\frac{n}{3}y} \sin(nx) + B_n e^{-\frac{n}{3}y} \sin(nx) \\ &= \left(1 - \frac{1}{e^{-\frac{2\pi}{3}} - 1}\right) e^{\frac{y}{3}} \sin(x) + \frac{1}{e^{-\frac{2\pi}{3}} - 1} e^{-\frac{y}{3}} \sin(x) \\ &= \left[\left(1 - \frac{1}{e^{-\frac{2\pi}{3}} - 1}\right) e^{\frac{y}{3}} + \frac{1}{e^{-\frac{2\pi}{3}} - 1} e^{-\frac{y}{3}}\right] \sin(x). \end{aligned}$$

We can check that this solves the PDE by plugging the solution into the equation and boundary conditions. \square

10.2.3 Exercises

Exercise 10.2.5: Classify each of the following second order differential equations as hyperbolic, parabolic, or elliptic.

a) $2\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial x \partial y} + 4\frac{\partial^2 u}{\partial y^2} + 3\frac{\partial u}{\partial x} - 5u = 0$

Exercise 10.2.6: Use separation variables to find a nontrivial solution to $u_{xx} + u_{yy} = 0$, where $u(x, 0) = 0$ and $u(0, y) = 0$. Hint: Try $u(x, y) = X(x)Y(y)$.

Exercise 10.2.7:* Use separation of variables to find a nontrivial solution to $u_{xt} = u_{xx}$.

Exercise 10.2.8:* Use separation of variables (Hint: try $u(x, t) = X(x) + T(t)$) to find a nontrivial solution to $u_x + u_t = u$.

TODO

Add more exercises. Look at what people normally assign.

10.3 The heat equation

Attribution: [JL], §4.6.

Learning Objectives

After this section, you will be able to:

- Use the heat equation to model the temperature in an object over time and
- Use separation of variables and Fourier series to solve the heat equation in specific domains.

10.3.1 Derivation of the heat equation

The first type of physical situation we want to consider is how the temperature of an object changes over time. We have dealt with this idea previously for a well-mixed fluid at a single temperature with Newton's Law of cooling. However, we will consider a physical object here, so the temperature will depend on position as well, giving rise to a PDE instead of the ODE we got with Newton's Law.

Let's start by considering an insulated wire of length L . With this, we mean that no heat can escape through the lateral sides of the wire, and the only way that heat can enter the system is through the two ends. We will let $u(t, x)$ be the temperature in the wire after t seconds, and at the position x meters from the left end of the wire (the units here are not important, they just need to be consistent). See Figure 10.5.

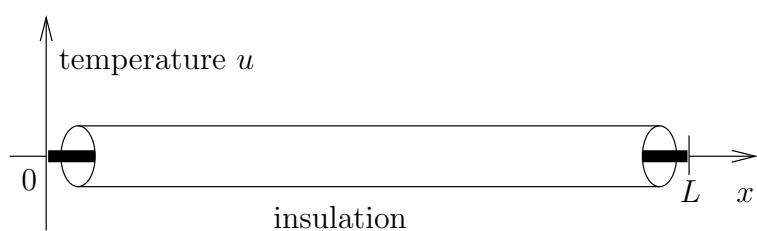


Figure 10.5: Insulated wire.

We want to come up with a way to find a differential equation that models how this temperature will change with time and location. The easiest way to do this is with the accumulation equation from § 1.9 with thermal energy, since that is something that can be accumulated, while temperature really can't. How does temperature relate to thermal energy? Physical properties tell us that

$$\Delta E = c\rho V \delta T$$

where ΔE is the change (or input) of energy, c is the *specific heat* of the material, ρ is the mass density of the material, V is the volume, and δT is the change in temperature. This

expression mainly comes from the fact that the specific heat is the energy required to raise one kilogram of the material by one degree Celsius.

Now, we need to build up the accumulation equation for energy. To do this, we will look at a small interior segment of this wire of length Δx that stretches from a position x to $x + \Delta x$. The accumulation equation tells us that

$$\text{rate of change of energy} = \text{rate in from left} + \text{rate in from right} + \text{interior sources}$$

For interior sources, we will assume that there is some function $Q(t, x)$ that describes all of the sources of thermal energy per unit volume at time t and position x , and our discussion of the relationship between temperature and thermal energy says that the rate of change of energy is

$$\frac{\partial E}{\partial t} = c\rho(A\Delta x)\frac{\partial u}{\partial t}$$

since u represents our temperature and $A\Delta x$ is the volume of this small piece of wire. Since the piece is “small”, we will assume that things like the energy change are constant along this piece, which works out using linearization since we are eventually going to limit Δx to zero.

At this point, we have the equation

$$c(x)\rho(x)(A\Delta x)\frac{\partial u}{\partial t} = \text{rate in from left} + \text{rate in from right} + Q(t, x)(A\Delta x)$$

where we have also added in the fact that c and ρ , these material properties, could depend on x if the wire is not uniform and made of the same material throughout. The last things we have to deal with are these rate terms. The normal terminology used for the rate of energy entering or exiting a region is *flux* of energy. Fourier’s law (same Fourier) says that if ϕ is the flux of energy, then

$$\phi = -k\nabla u$$

where u represents temperature and k is the *thermal conductivity* of the material. The point of this expression for now is the fact that the thermal flux is proportional to the derivative of temperature. This means that in this particular case

$$\text{flux in from left} = -k(x)A\frac{\partial u}{\partial x}(t, x)$$

and

$$\text{flux in from right} = k(x + \Delta x)A\frac{\partial u}{\partial x}(t, x + \Delta x).$$

The minus sign is missing from the second term because the flux coming in from the right is moving in the negative direction, so it needs to pick up an additional minus sign. The A here represents the cross-sectional area of the wire. Thus, we have the full equation

$$c(x)\rho(x)(A\Delta x)\frac{\partial u}{\partial t} = -k(x)A\frac{\partial u}{\partial x}(t, x) + k(x + \Delta x)A\frac{\partial u}{\partial x}(t, x + \Delta x) + Q(t, x)(A\Delta x).$$

Dividing both sides by $A\Delta x$ gives

$$c(x)\rho(x)\frac{\partial u}{\partial t} = \frac{1}{\Delta x} \left(k(x + \Delta x)\frac{\partial u}{\partial x}(t, x + \Delta x) - k(x)\frac{\partial u}{\partial x}(t, x) \right) + Q(t, x).$$

Now, we want to take the limit as $\Delta x \rightarrow 0$. The one term that still involves Δx looks like

$$\frac{k(x + \Delta x) \frac{\partial u}{\partial x}(t, x + \Delta x) - k(x) \frac{\partial u}{\partial x}(t, x)}{\Delta x}$$

which, as $\Delta x \rightarrow 0$ is a partial derivative. Taking this limit, we get the equation

$$c(x)\rho(x) \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left(k(x) \frac{\partial u}{\partial x} \right) + Q(t, x).$$

In most cases, we will assume that the material we use is uniform. That means that c , ρ and k are all independent of x . We can move the k outside the derivative, then divide both sides by $c\rho$ to get the equation

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} + F(t, x)$$

where

$$\alpha = \frac{k}{c\rho}$$

is the *thermal diffusivity* of the material and

$$F(t, x) = \frac{Q(t, x)}{c\rho}$$

represents the thermal sources in the system. This is the *heat equation*, in particular, this is a non-homogeneous heat equation. The homogeneous version is

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2}.$$

For an interpretation, this tells us that the change in heat at a specific point is proportional to the second derivative of the heat along the wire. This makes sense; if at a fixed t the graph of the heat distribution has a maximum (the graph is concave down), then heat flows away from the maximum. And vice versa.

Boundary conditions for the heat equation

There are several main types of boundary conditions that can be used with the heat equation. In general, we will assume that either the ends of the wire are exposed and fixed at a constant temperature, or they are insulated. In the first case, the endpoint being fixed at a constant temperature means that

$$u(t, 0) = T_0$$

at the left endpoint and

$$u(t, L) = T_L$$

at the right endpoint. For insulated ends, this means that thermal energy can not escape from the endpoint. Since the flux is proportional to the derivative of temperature, this means that

$$\frac{\partial u}{\partial x}(t, 0) = 0$$

at the left endpoint and

$$\frac{\partial u}{\partial x}(t, L) = 0$$

at the right endpoint. When trying to solve these problems, we want to have conditions where something is set to zero at the end point; that will make our lives much easier once we separate variables. These side conditions are said to be *homogeneous* (i.e., u or a derivative of u is set to zero).

It is possible to have more involved boundary conditions that involve both u and the derivative. These are sometimes more physically natural, since they are similar to Newton's law of cooling, but are much harder to solve. These may look something like

$$\frac{\partial u}{\partial x}(t, 0) - 4u(t, 0) = 0,$$

but we will generally avoid these types of conditions here.

We also need an initial condition—the temperature distribution at time $t = 0$. That is,

$$u(x, 0) = f(x),$$

for some known function $f(x)$. This initial condition is not a homogeneous side condition.

Multidimensional heat equation

The same ideas can be used to derive the heat equation in plates (two dimensions) or regions (three dimensions) as opposed to just one dimensional wires. The process is the same: take a small region and relate the change of energy to the flux through each side. In the two dimensional case, there are four sides to consider, two where the flux is related to $\frac{\partial u}{\partial x}$ and two where it is related to $\frac{\partial u}{\partial y}$. The differences end up working out the same way and with the same assumptions as before, this results in the heat equation in two dimensions looking like

$$\frac{\partial u}{\partial t} = \alpha \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + F(t, x, y).$$

With three spatial dimensions, there are six different flux terms, and the equation simplifies to

$$\frac{\partial u}{\partial t} = \alpha \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) + F(t, x, y, z).$$

In each case, we end up with a term that is the sum of all of the different second partial derivatives in the spatial directions. This type of a term shows up very frequently, so we give it a name.

Definition 10.3.1

Let u be a multivariable function. The *Laplacian* of u , denoted Δu is the sum of all second partial derivatives of u in the spatial directions. It does not take the time dependence of u into account, if u depends on time. Examples:

- One space dimension - $u(x)$ or $u(t, x)$: $\Delta u = \frac{\partial^2 u}{\partial x^2}$
- Two space dimensions - $u(x, y)$ or $u(t, x, y)$: $\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$
- Three space dimensions - $u(x, y, z)$ or $u(t, x, y, z)$: $\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}$.

With all of these definitions, the homogeneous heat equation, in any number of dimensions, can be written as

$$\frac{\partial u}{\partial t} = \alpha \Delta u.$$

Boundary conditions in the multidimensional case work similarly to the one-dimensional case. The temperature can either be fixed at the boundary, giving a value for the function u , or insulated, giving that the flux must be zero, or that

$$\nabla u \cdot \vec{n} = 0$$

along the boundary of the domain. For a rectangular domain, this will look like

$$\frac{\partial u}{\partial x} = 0 \quad \text{or} \quad \frac{\partial u}{\partial y} = 0.$$

10.3.2 Solution by separation of variables

Now, we want to employ techniques from § 10.2 to try to solve the heat equation for a variety of initial and boundary conditions.

Example 10.3.1: Let us try to solve the heat equation with fixed temperature endpoints:

$$u_t = ku_{xx} \quad \text{with} \quad u(0, t) = 0, \quad u(L, t) = 0, \quad \text{and} \quad u(x, 0) = f(x).$$

Solution: We use the method of separation of variables and guess $u(x, t) = X(x)T(t)$. We will try to make this guess satisfy the differential equation, $u_t = ku_{xx}$, and the homogeneous side conditions, $u(0, t) = 0$ and $u(L, t) = 0$. Then, as superposition preserves the differential equation and the homogeneous side conditions, we will try to build up a solution from these building blocks to solve the nonhomogeneous initial condition $u(x, 0) = f(x)$.

First we plug $u(x, t) = X(x)T(t)$ into the heat equation to obtain

$$X(x)T'(t) = kX''(x)T(t).$$

We rewrite as

$$\frac{T'(t)}{kT(t)} = \frac{X''(x)}{X(x)},$$

which means that both sides of this equation must equal a constant value $-\lambda$.

We obtain the two equations

$$\frac{T'(t)}{kT(t)} = -\lambda = \frac{X''(x)}{X(x)}.$$

In other words,

$$\begin{aligned} X''(x) + \lambda X(x) &= 0, \\ T'(t) + \lambda kT(t) &= 0. \end{aligned}$$

The boundary condition $u(0, t) = 0$ implies $X(0)T(t) = 0$. We are looking for a nontrivial solution and so we can assume that $T(t)$ is not identically zero. Hence $X(0) = 0$. Similarly, $u(L, t) = 0$ implies $X(L) = 0$. We are looking for nontrivial solutions X of the eigenvalue problem $X'' + \lambda X = 0$, $X(0) = 0$, $X(L) = 0$. We have previously found that the only eigenvalues are $\lambda_n = \frac{n^2\pi^2}{L^2}$, for integers $n \geq 1$, where eigenfunctions are $\sin(\frac{n\pi}{L}x)$. Hence, let us pick the solutions

$$X_n(x) = \sin\left(\frac{n\pi}{L}x\right).$$

The corresponding T_n must satisfy the equation

$$T'_n(t) + \frac{n^2\pi^2}{L^2}kT_n(t) = 0.$$

This is one of our **fundamental equations**, and the solution is just an exponential:

$$T_n(t) = e^{\frac{-n^2\pi^2}{L^2}kt}.$$

It will be useful to note that $T_n(0) = 1$. Our building-block solutions are

$$u_n(x, t) = X_n(x)T_n(t) = \sin\left(\frac{n\pi}{L}x\right)e^{\frac{-n^2\pi^2}{L^2}kt}.$$

We note that $u_n(x, 0) = \sin\left(\frac{n\pi}{L}x\right)$. Let us write $f(x)$ as the sine series

$$f(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi}{L}x\right).$$

That is, we find the Fourier series of the odd periodic extension of $f(x)$. We used the sine series as it corresponds to the eigenvalue problem for $X(x)$ above. Finally, we use superposition to write the solution as

$$u(x, t) = \sum_{n=1}^{\infty} b_n u_n(x, t) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi}{L}x\right) e^{\frac{-n^2\pi^2}{L^2}kt}.$$

□

Why does this solution work? First note that it is a solution to the heat equation by superposition. It satisfies $u(0, t) = 0$ and $u(L, t) = 0$, because $x = 0$ or $x = L$ makes all the sines vanish. Finally, plugging in $t = 0$, we notice that $T_n(0) = 1$ and so

$$u(x, 0) = \sum_{n=1}^{\infty} b_n u_n(x, 0) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi}{L}x\right) = f(x).$$

Example 10.3.2: Consider an insulated wire of length 1 whose ends are embedded in ice (temperature 0). Let $k = 0.003$ and let the initial heat distribution be $u(x, 0) = 50x(1-x)$. See Figure 10.6. Suppose we want to find the temperature function $u(x, t)$. Let us also suppose we want to find when (at what t) does the maximum temperature in the wire drop to one half of the initial maximum of 12.5.

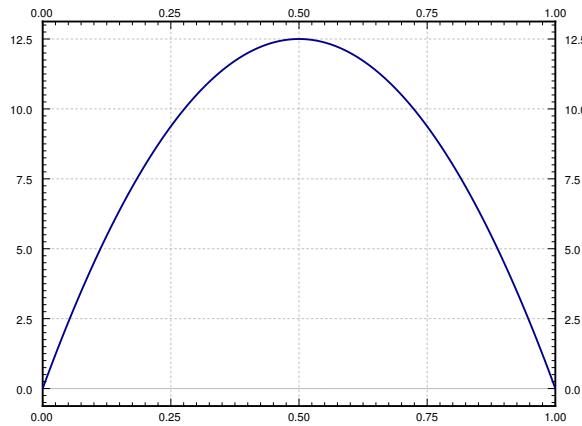


Figure 10.6: Initial distribution of temperature in the wire.

Solution: We are solving the following PDE problem:

$$\begin{aligned} u_t &= 0.003 u_{xx}, \\ u(0, t) &= u(1, t) = 0, \\ u(x, 0) &= 50x(1-x) \quad \text{for } 0 < x < 1. \end{aligned}$$

We write $f(x) = 50x(1-x)$ for $0 < x < 1$ as a sine series. That is, $f(x) = \sum_{n=1}^{\infty} b_n \sin(n\pi x)$, where

$$b_n = 2 \int_0^1 50x(1-x) \sin(n\pi x) dx = \frac{200}{\pi^3 n^3} - \frac{200(-1)^n}{\pi^3 n^3} = \begin{cases} 0 & \text{if } n \text{ even,} \\ \frac{400}{\pi^3 n^3} & \text{if } n \text{ odd.} \end{cases}$$

The solution $u(x, t)$, plotted in Figure 10.7 on the following page for $0 \leq t \leq 100$, is given by the series:

$$u(x, t) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{400}{\pi^3 n^3} \sin(n\pi x) e^{-n^2 \pi^2 0.003 t}.$$

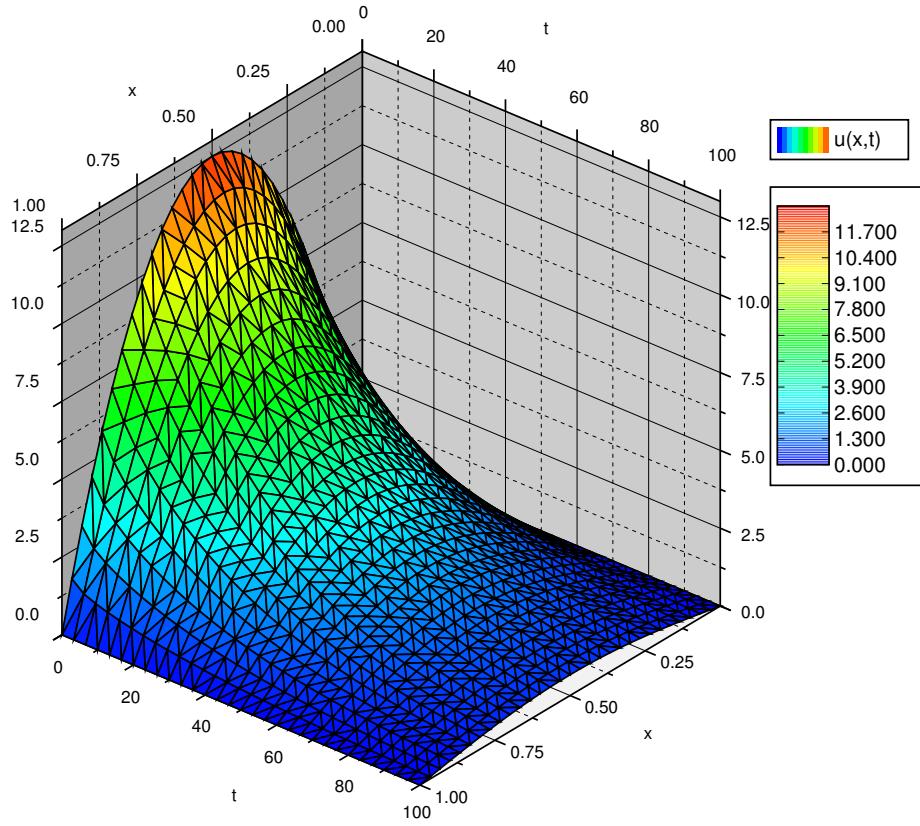


Figure 10.7: Plot of the temperature of the wire at position x at time t .

Finally, let us answer the question about the maximum temperature. It is relatively easy to see that the maximum temperature at any fixed time is always at $x = 0.5$, in the middle of the wire. The plot of $u(x, t)$ confirms this intuition. If we plug in $x = 0.5$, we get

$$u(0.5, t) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{400}{\pi^3 n^3} \sin(n\pi 0.5) e^{-n^2 \pi^2 0.003 t}.$$

For $n = 3$ and higher (remember n is only odd), the terms of the series are insignificant compared to the first term. The first term in the series is already a very good approximation of the function. Hence

$$u(0.5, t) \approx \frac{400}{\pi^3} e^{-\pi^2 0.003 t}.$$

The approximation gets better and better as t gets larger as the other terms decay much faster. Let us plot the function $u(0.5, t)$, the temperature at the midpoint of the wire at time t , in Figure 10.8 on the next page. The figure also plots the approximation by the first term.

After $t = 5$ or so it would be hard to tell the difference between the first term of the series for $u(x, t)$ and the real solution $u(x, t)$. This behavior is a general feature of solving the heat equation. If you are interested in behavior for large enough t , only the first one or two terms may be necessary.

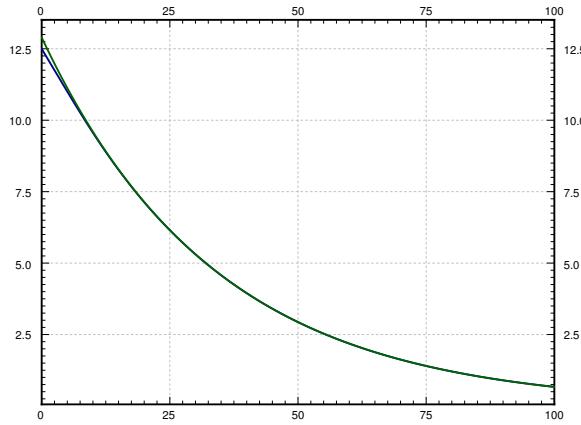


Figure 10.8: Temperature at the midpoint of the wire (the bottom curve), and the approximation of this temperature by using only the first term in the series (top curve).

Let us get back to the question of when is the maximum temperature one half of the initial maximum temperature. That is, when is the temperature at the midpoint $12.5/2 = 6.25$. We notice on the graph that if we use the approximation by the first term we will be close enough. We solve

$$6.25 = \frac{400}{\pi^3} e^{-\pi^2 0.003 t}.$$

That is,

$$t = \frac{\ln \frac{6.25 \pi^3}{400}}{-\pi^2 0.003} \approx 24.5.$$

So the maximum temperature drops to half at about $t = 24.5$. □

We mention an interesting behavior of the solution to the heat equation. The heat equation “smoothes” out the function $f(x)$ as t grows. For a fixed t , the solution is a Fourier series with coefficients $b_n e^{\frac{-n^2 \pi^2}{L^2} kt}$. If $t > 0$, then these coefficients go to zero faster than any $\frac{1}{n^p}$ for any power p . In other words, the Fourier series has infinitely many derivatives everywhere. Thus even if the function $f(x)$ has jumps and corners, then for a fixed $t > 0$, the solution $u(x, t)$ as a function of x is as smooth as we want it to be.

Example 10.3.3: When the initial condition is already a sine series, then there is no need to compute anything, you just need to plug in. Consider

$$u_t = 0.3 u_{xx}, \quad u(0, t) = u(1, t) = 0, \quad u(x, 0) = 0.1 \sin(\pi t) + \sin(2\pi t).$$

The solution is then

$$u(x, t) = 0.1 \sin(\pi t) e^{-0.3\pi^2 t} + \sin(2\pi t) e^{-1.2\pi^2 t}.$$

10.3.3 Insulated ends

Now suppose the ends of the wire are insulated. In this case, we are solving the equation

$$u_t = ku_{xx} \quad \text{with} \quad u_x(0, t) = 0, \quad u_x(L, t) = 0, \quad \text{and} \quad u(x, 0) = f(x).$$

Yet again we try a solution of the form $u(x, t) = X(x)T(t)$. By the same procedure as before we plug into the heat equation and arrive at the following two equations

$$\begin{aligned} X''(x) + \lambda X(x) &= 0, \\ T'(t) + \lambda k T(t) &= 0. \end{aligned}$$

At this point the story changes slightly. The boundary condition $u_x(0, t) = 0$ implies $X'(0)T(t) = 0$. Hence $X'(0) = 0$. Similarly, $u_x(L, t) = 0$ implies $X'(L) = 0$. We are looking for nontrivial solutions X of the eigenvalue problem $X'' + \lambda X = 0$, $X'(0) = 0$, $X'(L) = 0$. We have previously found that the only eigenvalues are $\lambda_n = \frac{n^2\pi^2}{L^2}$, for integers $n \geq 0$, where eigenfunctions are $\cos\left(\frac{n\pi}{L}x\right)$ (we include the constant eigenfunction). Hence, let us pick solutions

$$X_n(x) = \cos\left(\frac{n\pi}{L}x\right) \quad \text{and} \quad X_0(x) = 1.$$

The corresponding T_n must satisfy the equation

$$T'_n(t) + \frac{n^2\pi^2}{L^2}k T_n(t) = 0.$$

For $n \geq 1$, as before,

$$T_n(t) = e^{\frac{-n^2\pi^2}{L^2}kt}.$$

For $n = 0$, we have $T'_0(t) = 0$ and hence $T_0(t) = 1$. Our building-block solutions are

$$u_n(x, t) = X_n(x)T_n(t) = \cos\left(\frac{n\pi}{L}x\right)e^{\frac{-n^2\pi^2}{L^2}kt},$$

and

$$u_0(x, t) = 1.$$

We note that $u_n(x, 0) = \cos\left(\frac{n\pi}{L}x\right)$. Let us write f using the cosine series

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}x\right).$$

That is, we find the Fourier series of the even periodic extension of $f(x)$.

We use superposition to write the solution as

$$u(x, t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n u_n(x, t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}x\right)e^{\frac{-n^2\pi^2}{L^2}kt}.$$

Example 10.3.4: Let us try the same equation as before, but for insulated ends. We are solving the following PDE problem

$$\begin{aligned} u_t &= 0.003 u_{xx}, \\ u_x(0, t) &= u_x(1, t) = 0, \\ u(x, 0) &= 50x(1-x) \quad \text{for } 0 < x < 1. \end{aligned}$$

Solution: For this problem, we must find the cosine series of $u(x, 0)$. For $0 < x < 1$ we have

$$50x(1-x) = \frac{25}{3} + \sum_{\substack{n=2 \\ n \text{ even}}}^{\infty} \left(\frac{-200}{\pi^2 n^2} \right) \cos(n\pi x).$$

The calculation is left to the reader. Hence, the solution to the PDE problem, plotted in Figure 10.9, is given by the series

$$u(x, t) = \frac{25}{3} + \sum_{\substack{n=2 \\ n \text{ even}}}^{\infty} \left(\frac{-200}{\pi^2 n^2} \right) \cos(n\pi x) e^{-n^2 \pi^2 0.003 t}.$$

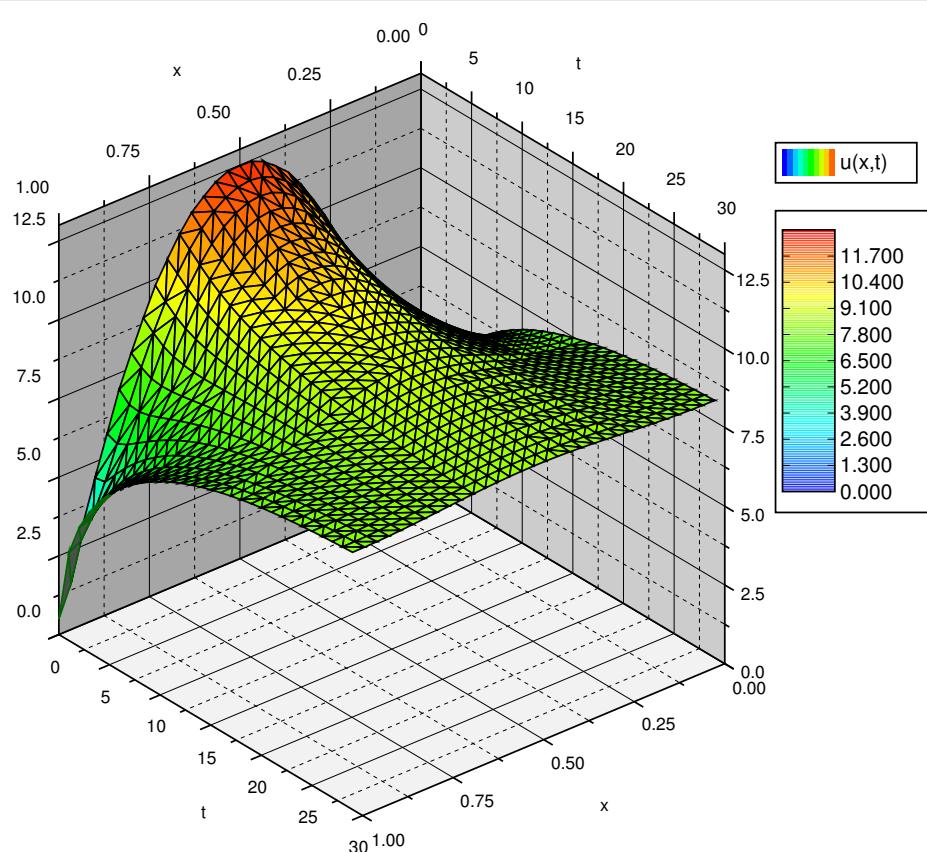


Figure 10.9: Plot of the temperature of the insulated wire at position x at time t .

Note in the graph that as time goes on, the temperature evens out across the wire. Eventually, all the terms except the constant die out, and you will be left with a uniform temperature of $\frac{25}{3} \approx 8.33$ along the entire length of the wire. \square

Let us expand on the last point. The constant term in the series is

$$\frac{a_0}{2} = \frac{1}{L} \int_0^L f(x) dx.$$

In other words, $\frac{a_0}{2}$ is the average value of $f(x)$, that is, the average of the initial temperature. As the wire is insulated everywhere, no heat can get out, no heat can get in. So the temperature tries to distribute evenly over time, and the average temperature must always be the same, in particular it is always $\frac{a_0}{2}$. As time goes to infinity, the temperature goes to the constant $\frac{a_0}{2}$ everywhere.

10.3.4 Exercises

Exercise 10.3.1: Consider a wire of length 2, with $k = 0.001$ and an initial temperature distribution $u(x, 0) = 50x$. Both ends are embedded in ice (temperature 0). Find the solution as a series.

Exercise 10.3.2: Find a series solution of

$$\begin{aligned} u_t &= u_{xx}, \\ u(0, t) &= u(1, t) = 0, \\ u(x, 0) &= 100 \quad \text{for } 0 < x < 1. \end{aligned}$$

Exercise 10.3.3:* Find a series solution of

$$\begin{aligned} u_t &= 3u_{xx}, \\ u(0, t) &= u(\pi, t) = 0, \\ u(x, 0) &= 5 \sin(x) + 2 \sin(5x) \quad \text{for } 0 < x < \pi. \end{aligned}$$

Exercise 10.3.4: Find a series solution of

$$\begin{aligned} u_t &= u_{xx}, \\ u_x(0, t) &= u_x(\pi, t) = 0, \\ u(x, 0) &= 3 \cos(x) + \cos(3x) \quad \text{for } 0 < x < \pi. \end{aligned}$$

Exercise 10.3.5:* Find a series solution of

$$\begin{aligned} u_t &= 0.1u_{xx}, \\ u_x(0, t) &= u_x(\pi, t) = 0, \\ u(x, 0) &= 1 + 2 \cos(x) \quad \text{for } 0 < x < \pi. \end{aligned}$$

Exercise 10.3.6: Find a series solution of

$$\begin{aligned} u_t &= \frac{1}{3}u_{xx}, \\ u_x(0, t) &= u_x(\pi, t) = 0, \\ u(x, 0) &= \frac{10x}{\pi} \quad \text{for } 0 < x < \pi. \end{aligned}$$

Exercise 10.3.7: Find a series solution of

$$\begin{aligned} u_t &= u_{xx}, \\ u(0, t) &= 0, \quad u(1, t) = 100, \\ u(x, 0) &= \sin(\pi x) \quad \text{for } 0 < x < 1. \end{aligned}$$

Hint: Use the fact that $u(x, t) = 100x$ is a solution satisfying $u_t = u_{xx}$, $u(0, t) = 0$, $u(1, t) = 100$. Then use superposition.

Exercise 10.3.8: Find the steady state temperature solution as a function of x alone, by letting $t \rightarrow \infty$ in the solution from exercises 10.3.6 and 10.3.7. Verify that it satisfies the equation $u_{xx} = 0$.

Exercise 10.3.9 (challenging): Suppose that one end of the wire is insulated (say at $x = 0$) and the other end is kept at zero temperature. That is, find a series solution of

$$\begin{aligned} u_t &= ku_{xx}, \\ u_x(0, t) &= u(L, t) = 0, \\ u(x, 0) &= f(x) \quad \text{for } 0 < x < L. \end{aligned}$$

Express any coefficients in the series by integrals of $f(x)$.

Exercise 10.3.10 (challenging): Suppose that the wire is circular and insulated, so there are no ends. You can think of this as simply connecting the two ends and making sure the solution matches up at the ends. That is, find a series solution of

$$\begin{aligned} u_t &= ku_{xx}, \\ u(0, t) &= u(L, t), \quad u_x(0, t) = u_x(L, t), \\ u(x, 0) &= f(x) \quad \text{for } 0 < x < L. \end{aligned}$$

Express any coefficients in the series by integrals of $f(x)$.

Exercise 10.3.11: Consider a wire insulated on both ends, $L = 1$, $k = 1$, and $u(x, 0) = \cos^2(\pi x)$.

- a) Find the solution $u(x, t)$. Hint: a trig identity.
- b) Find the average temperature.
- c) Initially the temperature variation is 1 (maximum minus the minimum). Find the time when the variation is $1/2$.

Exercise 10.3.12:* Suppose that the temperature on the wire is fixed at 0 at the ends, $L = 1$, $k = 1$, and $u(x, 0) = 100 \sin(2\pi x)$.

- a) What is the temperature at $x = 1/2$ at any time.
- b) What is the maximum and the minimum temperature on the wire at $t = 0$.
- c) At what time is the maximum temperature on the wire exactly one half of the initial maximum at $t = 0$.

10.4 One-dimensional wave equation

Attribution: [JL], §4.7.

Learning Objectives

After this section, you will be able to:

- Identify the one-dimensional wave equation and
- Use superposition and Fourier series to solve the equation using separation of variables.

Imagine we have a tensioned guitar string of length L . Let us only consider vibrations in one direction. Let x denote the position along the string, let t denote time, and let y denote the displacement of the string from the rest position. See Figure 10.10.

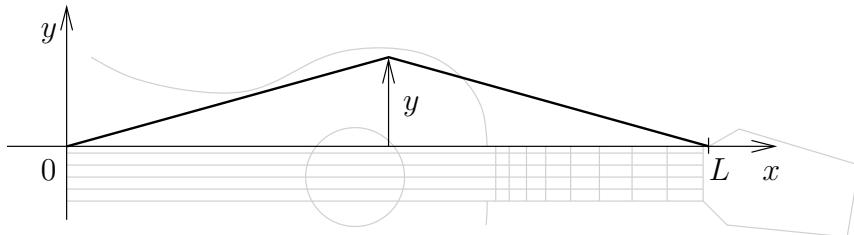


Figure 10.10: Vibrating string of length L , x is position, y is displacement.

10.4.1 Derivation of the wave equation

We want to determine a partial differential equation that governs the displacement of this string over time. Similar to what we did with the heat equation, we will consider a small section of the string of length Δx , from x to $x + \Delta x$.

The basis for the equation that we will develop here is Newton's second law, that force equals mass times acceleration. The diagram in Figure 10.11 shows the forces acting on this small section of string. For the mass, we will assume that ρ is the mass of the string per unit length, and will multiply this by the length of the string to get mass. There are internal forces $F(t, x)$ that are applied in the transverse direction as well as tension forces T on each end of the string. The tension at point (t, x) has magnitude $T(t, x)$ and pulls at angle $\theta(t, x)$ from the horizontal. We can use this to build two equations involving this segment of string, one for the forces in the vertical (transverse) direction, and one in the horizontal direction (along the string).

In the horizontal direction, we will assume that the displacement is always zero; that is, the string does not move in the horizontal direction as it vibrates. Therefore, this equation, taking into account all forces in the horizontal direction, gives

$$0 = T(t, x + \Delta x) \cos(\theta(t, x + \Delta x)) - T(t, x) \cos(\theta(t, x))$$

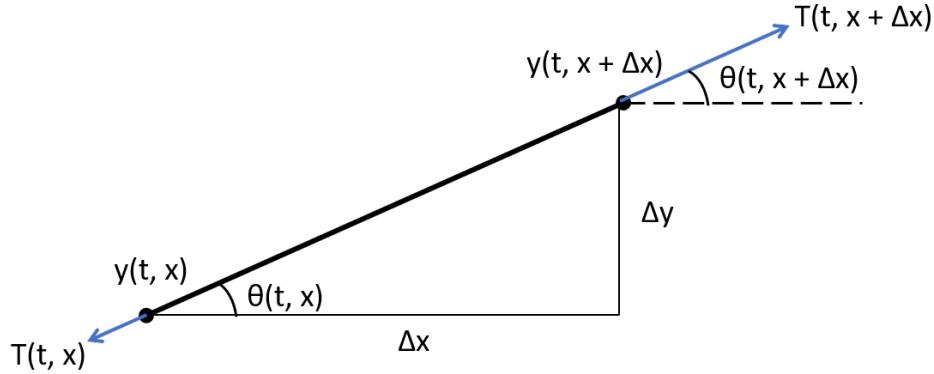


Figure 10.11: Forces acting on a small portion of the spring.

since based on the diagram, $T \cos(\theta)$ is the horizontal component of the tension force at each endpoint. In the vertical direction, we get

$$\rho(x) \sqrt{(\Delta x)^2 + (\Delta y)^2} \frac{\partial^2 y}{\partial t^2} = F(t, x)(\Delta x) + T(t, x + \Delta x) \sin(\theta(t, x + \Delta x)) - T(t, x) \sin(\theta(t, x)).$$

As with the heat equation work, we are going to divide both sides of each equation by Δx and then take the limit as $\Delta x \rightarrow 0$. With this limit, we have that

$$\lim_{\Delta x \rightarrow 0} T(t, x + \Delta x) \cos(\theta(t, x + \Delta x)) - T(t, x) \cos(\theta(t, x)) = \frac{\partial}{\partial x} (T(t, x) \cos(\theta(t, x)))$$

and

$$\lim_{\Delta x \rightarrow 0} T(t, x + \Delta x) \sin(\theta(t, x + \Delta x)) - T(t, x) \sin(\theta(t, x)) = \frac{\partial}{\partial x} (T(t, x) \sin(\theta(t, x))).$$

In addition, the left side of the vertical component equation, we can simplify it to

$$\begin{aligned} \lim_{\Delta x \rightarrow 0} \frac{1}{\Delta x} \sqrt{(\Delta x)^2 + (\Delta y)^2} &= \lim_{\Delta x \rightarrow 0} \sqrt{1 + \frac{(\Delta y)^2}{(\Delta x)^2}} \\ &= \lim_{\Delta x \rightarrow 0} \sqrt{1 + \left(\frac{\Delta y}{\Delta x}\right)^2} \\ &= \sqrt{1 + \left(\frac{\partial y}{\partial x}\right)^2} \end{aligned}$$

Thus, the horizontal component of the force balance becomes

$$0 = \frac{\partial}{\partial x} (T(t, x) \cos(\theta(t, x)))$$

and the vertical component is

$$\rho(x) \sqrt{1 + \left(\frac{\partial y}{\partial x}\right)^2} \frac{\partial^2 y}{\partial t^2} = F(t, x) + \frac{\partial}{\partial x} (T(t, x) \sin(\theta(t, x))).$$

We want to start by simplifying the vertical equation, which we can first do by using the product rule to get

$$\rho(x) \sqrt{1 + \left(\frac{\partial y}{\partial x}\right)^2} \frac{\partial^2 y}{\partial t^2} = F(t, x) + \frac{\partial T}{\partial x} \sin(\theta(t, x)) + T(t, x) \cos(\theta(t, x)) \frac{\partial \theta}{\partial x}. \quad (10.3)$$

The first thing we want to do is convert the θ variables to be written in terms of y . Based on the triangle in [Figure 10.11](#), we can see that

$$\tan(\theta) = \frac{\Delta y}{\Delta x}$$

so that in the limit as $\Delta x \rightarrow 0$, $\tan(\theta) = \frac{\partial y}{\partial x}$. By drawing another triangle, we can find that

$$\sin(\theta) = \frac{\frac{\partial y}{\partial x}}{\sqrt{1 + \left(\frac{\partial y}{\partial x}\right)^2}} \quad \cos(\theta) = \frac{1}{\sqrt{1 + \left(\frac{\partial y}{\partial x}\right)^2}}$$

as well as

$$\theta = \tan^{-1}\left(\frac{\partial y}{\partial x}\right) \quad \frac{\partial \theta}{\partial x} = \frac{\frac{\partial^2 y}{\partial x^2}}{1 + \left(\frac{\partial y}{\partial x}\right)^2}.$$

We could plug all of this into (10.3), but that would be really messy. However, to get a simpler equation, we will assume that we are looking at “small” vibrations, which will make things easier (and linear). This means that we assume that

$$\sqrt{1 + \left(\frac{\partial y}{\partial x}\right)^2} \approx 1$$

so that

$$\sin(\theta) \approx \frac{\partial y}{\partial x} \quad \cos(\theta) \approx 1 \quad \frac{\partial \theta}{\partial x} \approx \frac{\partial^2 y}{\partial x^2}.$$

Plugging these into (10.3) gives

$$\rho(x) \frac{\partial^2 y}{\partial t^2} = F(t, x) + \frac{\partial T}{\partial x} \frac{\partial y}{\partial x} + T(t, x) \frac{\partial^2 y}{\partial x^2}. \quad (10.4)$$

Next, we plug these approximations into the horizontal component after applying the product rule to get

$$\begin{aligned} 0 &= \frac{\partial}{\partial x} (T(t, x) \cos(\theta(t, x))) \\ &= \frac{\partial T}{\partial x} \cos(\theta(t, x)) + T(t, x) \sin(\theta(t, x)) \frac{\partial \theta}{\partial x} \\ &= \frac{\partial T}{\partial x} + T(t, x) \frac{\partial y}{\partial x} \frac{\partial^2 y}{\partial x^2} \\ &\approx \frac{\partial T}{\partial x} \end{aligned}$$

because $\frac{\partial y}{\partial x}$ and $\frac{\partial^2 y}{\partial x^2}$ are both small, so we can approximate the product by zero. Thus, we have that

$$\frac{\partial T}{\partial x} = 0$$

which also makes sense from a physical point of view, because it says that the tension in the string is constant throughout, so that it uniformly distributes across the string if you only pull on the two ends.

Putting this fact into (10.4) gives our final equation as

$$\rho(x) \frac{\partial^2 y}{\partial t^2} = F(t, x) + T(t, x) \frac{\partial^2 y}{\partial x^2}$$

In general, we will assume that the density ρ is independent of position (string is made of consistent material) and T is independent of time to give the equation

$$\frac{\partial^2 y}{\partial t^2} = G(t, x) + a \frac{\partial^2 y}{\partial x^2} \quad (10.5)$$

where a is the constant $\frac{T}{\rho}$ and G is a scaled version of the external forces. If $G = 0$, we get the homogeneous wave equation.

This equation can also be written in multiple spacial dimensions. As you might guess, the second derivative term becomes the Laplacian when we add other spacial dimensions. Thus, if y is the displacement, the homogeneous wave equation is that

$$\frac{\partial^2 y}{\partial t^2} = a \Delta y.$$

10.4.2 Boundary and initial conditions

As derived previously, the *one-dimensional wave equation* is:

$$y_{tt} = a^2 y_{xx},$$

for some constant $a > 0$. The intuition is similar to the heat equation, replacing velocity with acceleration: the acceleration at a specific point is proportional to the second derivative of the shape of the string. In other words when the string is concave down then u_{xx} is negative and the string wants to accelerate downwards, so u_{tt} should be negative. And vice versa. The wave equation is an example of a hyperbolic PDE.

Assume that the ends of the string are fixed in place as on the guitar:

$$y(0, t) = 0 \quad \text{and} \quad y(L, t) = 0.$$

Note that we have two conditions along the x -axis as there are two derivatives in the x direction.

There are also two derivatives along the t direction and hence we need two further conditions here. We need to know the initial position and the initial velocity of the string. That is, for some known functions $f(x)$ and $g(x)$, we impose

$$y(x, 0) = f(x) \quad \text{and} \quad y_t(x, 0) = g(x).$$

The equation is linear, so superposition works just as it did for the heat equation. And again we will use separation of variables to find enough building-block solutions to get the overall solution. There is one change however. It will be easier to solve two separate problems and add their solutions.

The two problems we will solve are

$$\begin{aligned} w_{tt} &= a^2 w_{xx}, \\ w(0, t) &= w(L, t) = 0, \\ w(x, 0) &= 0 && \text{for } 0 < x < L, \\ w_t(x, 0) &= g(x) && \text{for } 0 < x < L, \end{aligned} \tag{10.6}$$

and

$$\begin{aligned} z_{tt} &= a^2 z_{xx}, \\ z(0, t) &= z(L, t) = 0, \\ z(x, 0) &= f(x) && \text{for } 0 < x < L, \\ z_t(x, 0) &= 0 && \text{for } 0 < x < L. \end{aligned} \tag{10.7}$$

The principle of superposition implies that $y = w + z$ solves the wave equation and furthermore $y(x, 0) = w(x, 0) + z(x, 0) = f(x)$ and $y_t(x, 0) = w_t(x, 0) + z_t(x, 0) = g(x)$. Hence, y is a solution to

$$\begin{aligned} y_{tt} &= a^2 y_{xx}, \\ y(0, t) &= y(L, t) = 0, \\ y(x, 0) &= f(x) && \text{for } 0 < x < L, \\ y_t(x, 0) &= g(x) && \text{for } 0 < x < L. \end{aligned} \tag{10.8}$$

The reason for all this complexity is that superposition only works for homogeneous conditions such as $y(0, t) = y(L, t) = 0$, $y(x, 0) = 0$, or $y_t(x, 0) = 0$. Therefore, we can use separation of variables to find many building-block solutions solving all the homogeneous conditions. We can then use them to construct a solution satisfying the remaining nonhomogeneous condition.

Let us start with (10.6). We try a solution of the form $w(x, t) = X(x)T(t)$ again. We plug into the wave equation to obtain

$$X(x)T''(t) = a^2 X''(x)T(t).$$

Rewriting we get

$$\frac{T''(t)}{a^2 T(t)} = \frac{X''(x)}{X(x)}.$$

Again, left-hand side depends only on t and the right-hand side depends only on x . So both sides equal a constant, which we denote by $-\lambda$:

$$\frac{T''(t)}{a^2 T(t)} = -\lambda = \frac{X''(x)}{X(x)}.$$

We solve to get two ordinary differential equations

$$\begin{aligned} X''(x) + \lambda X(x) &= 0, \\ T''(t) + \lambda a^2 T(t) &= 0. \end{aligned}$$

The conditions $0 = w(0, t) = X(0)T(t)$ implies $X(0) = 0$ and $w(L, t) = 0$ implies that $X(L) = 0$. Therefore, the only nontrivial solutions for the first equation are when $\lambda = \lambda_n = \frac{n^2\pi^2}{L^2}$ and they are

$$X_n(x) = \sin\left(\frac{n\pi}{L}x\right).$$

The general solution for T for this particular λ_n is

$$T_n(t) = A \cos\left(\frac{n\pi a}{L}t\right) + B \sin\left(\frac{n\pi a}{L}t\right).$$

We also have the condition that $w(x, 0) = 0$ or $X(x)T(0) = 0$. This implies that $T(0) = 0$, which in turn forces $A = 0$. It is convenient to pick $B = \frac{L}{n\pi a}$ (you will see why in a moment) and hence

$$T_n(t) = \frac{L}{n\pi a} \sin\left(\frac{n\pi a}{L}t\right).$$

Our building-block solutions are

$$w_n(x, t) = \frac{L}{n\pi a} \sin\left(\frac{n\pi}{L}x\right) \sin\left(\frac{n\pi a}{L}t\right).$$

We differentiate in t :

$$\frac{\partial w_n}{\partial t}(x, t) = \sin\left(\frac{n\pi}{L}x\right) \cos\left(\frac{n\pi a}{L}t\right).$$

Hence,

$$\frac{\partial w_n}{\partial t}(x, 0) = \sin\left(\frac{n\pi}{L}x\right).$$

We expand $g(x)$ in terms of these sines as

$$g(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi}{L}x\right).$$

Using superposition we write the solution to (10.6) as a series

$$w(x, t) = \sum_{n=1}^{\infty} b_n w_n(x, t) = \sum_{n=1}^{\infty} b_n \frac{L}{n\pi a} \sin\left(\frac{n\pi}{L}x\right) \sin\left(\frac{n\pi a}{L}t\right).$$

Exercise 10.4.1: Check that $w(x, 0) = 0$ and $w_t(x, 0) = g(x)$.

We solve (10.7) similarly. We again try $z(x, y) = X(x)T(t)$. The procedure works exactly the same at first. We obtain

$$\begin{aligned} X''(x) + \lambda X(x) &= 0, \\ T''(t) + \lambda a^2 T(t) &= 0, \end{aligned}$$

and the conditions $X(0) = 0$, $X(L) = 0$. So again $\lambda = \lambda_n = \frac{n^2\pi^2}{L^2}$ and

$$X_n(x) = \sin\left(\frac{n\pi}{L}x\right).$$

This time the condition on T is $T'(0) = 0$. Thus we get that $B = 0$ and we take

$$T_n(t) = \cos\left(\frac{n\pi a}{L}t\right).$$

Our building-block solution is

$$z_n(x, t) = \sin\left(\frac{n\pi}{L}x\right) \cos\left(\frac{n\pi a}{L}t\right).$$

As $z_n(x, 0) = \sin\left(\frac{n\pi}{L}x\right)$, we expand $f(x)$ in terms of these sines as

$$f(x) = \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi}{L}x\right).$$

And we write down the solution to (10.7) as a series

$$z(x, t) = \sum_{n=1}^{\infty} c_n z_n(x, t) = \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi}{L}x\right) \cos\left(\frac{n\pi a}{L}t\right).$$

Exercise 10.4.2: Fill in the details in the derivation of the solution of (10.7). Check that the solution satisfies all the side conditions.

Putting these two solutions together, let us state the result as a theorem.

Theorem 10.4.1

Take the equation

$$\begin{aligned} y_{tt} &= a^2 y_{xx}, \\ y(0, t) &= y(L, t) = 0, \\ y(x, 0) &= f(x) && \text{for } 0 < x < L, \\ y_t(x, 0) &= g(x) && \text{for } 0 < x < L, \end{aligned} \tag{10.9}$$

where

$$f(x) = \sum_{n=1}^{\infty} c_n \sin\left(\frac{n\pi}{L}x\right) \quad \text{and} \quad g(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi}{L}x\right).$$

Then the solution $y(x, t)$ can be written as a sum of the solutions of (10.6) and (10.7):

$$\begin{aligned} y(x, t) &= \sum_{n=1}^{\infty} b_n \frac{L}{n\pi a} \sin\left(\frac{n\pi}{L}x\right) \sin\left(\frac{n\pi a}{L}t\right) + c_n \sin\left(\frac{n\pi}{L}x\right) \cos\left(\frac{n\pi a}{L}t\right) \\ &= \sum_{n=1}^{\infty} \sin\left(\frac{n\pi}{L}x\right) \left[b_n \frac{L}{n\pi a} \sin\left(\frac{n\pi a}{L}t\right) + c_n \cos\left(\frac{n\pi a}{L}t\right) \right]. \end{aligned}$$

Example 10.4.1: Consider a string of length 2 plucked in the middle, it has an initial shape given in Figure 10.12 on the facing page. That is,

$$f(x) = \begin{cases} 0.1x & \text{if } 0 \leq x \leq 1, \\ 0.1(2-x) & \text{if } 1 < x \leq 2. \end{cases}$$

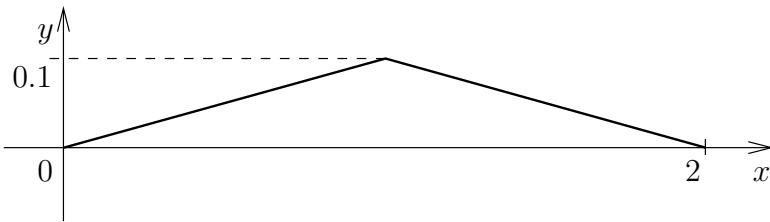


Figure 10.12: Initial shape of a plucked string from Example 10.4.1.

Let the string start at rest ($g(x) = 0$), and let $a = 1$ for simplicity. In other words, we wish to solve the problem:

$$\begin{aligned}y_{tt} &= y_{xx}, \\y(0, t) &= y(2, t) = 0, \\y(x, 0) &= f(x) \quad \text{and} \quad y_t(x, 0) = 0.\end{aligned}$$

We leave it to the reader to compute the sine series of $f(x)$. The series will be

$$f(x) = \sum_{n=1}^{\infty} \frac{0.8}{n^2\pi^2} \sin\left(\frac{n\pi}{2}\right) \sin\left(\frac{n\pi}{2}x\right).$$

Note that $\sin\left(\frac{n\pi}{2}\right)$ is the sequence $1, 0, -1, 0, 1, 0, -1, \dots$ for $n = 1, 2, 3, 4, \dots$. Therefore,

$$f(x) = \frac{0.8}{\pi^2} \sin\left(\frac{\pi}{2}x\right) - \frac{0.8}{9\pi^2} \sin\left(\frac{3\pi}{2}x\right) + \frac{0.8}{25\pi^2} \sin\left(\frac{5\pi}{2}x\right) - \dots$$

The solution $y(x, t)$ is given by

$$\begin{aligned}y(x, t) &= \sum_{n=1}^{\infty} \frac{0.8}{n^2\pi^2} \sin\left(\frac{n\pi}{2}\right) \sin\left(\frac{n\pi}{2}x\right) \cos\left(\frac{n\pi}{2}t\right) \\&= \sum_{m=1}^{\infty} \frac{0.8(-1)^{m+1}}{(2m-1)^2\pi^2} \sin\left(\frac{(2m-1)\pi}{2}x\right) \cos\left(\frac{(2m-1)\pi}{2}t\right) \\&= \frac{0.8}{\pi^2} \sin\left(\frac{\pi}{2}x\right) \cos\left(\frac{\pi}{2}t\right) - \frac{0.8}{9\pi^2} \sin\left(\frac{3\pi}{2}x\right) \cos\left(\frac{3\pi}{2}t\right) \\&\quad + \frac{0.8}{25\pi^2} \sin\left(\frac{5\pi}{2}x\right) \cos\left(\frac{5\pi}{2}t\right) - \dots\end{aligned}$$

See Figure 10.13 on the next page for a plot for $0 < t < 3$. Notice that unlike the heat equation, the solution does not become “smoother,” the “sharp edges” remain. We will see the reason for this behavior in the next section where we derive the solution to the wave equation in a different way.

Make sure you understand what the plot, such as the one in the figure, is telling you. For each fixed t , you can think of the function $y(x, t)$ as just a function of x . This function gives

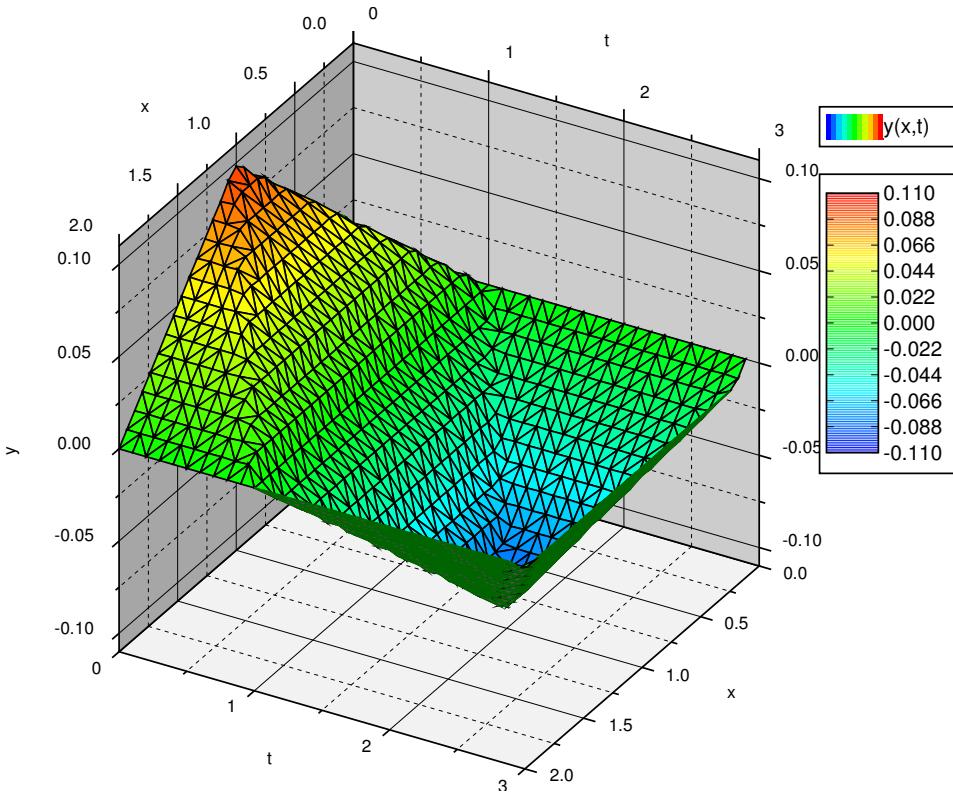


Figure 10.13: Shape of the plucked string for $0 < t < 3$.

you the shape of the string at time t . See Figure 10.14 on the facing page for plots of y as a function of x at several different values of t . On this plot you can see the sharp edges remaining much better.

One thing to take away from all this is how a guitar sounds. Notice that the (angular) frequencies that come up in the solution are $n\frac{\pi a}{L}$. That is, there is a certain base *fundamental frequency* $\frac{\pi a}{L}$, and then we also get all the multiples of this frequency, which in music are called the *overtones*. Which overtones appear and with what amplitude is what musicians call the *timbre* of the note. Mathematicians usually call this the *spectrum*. Because all the frequencies are multiples of one frequency (the fundamental) we get a nice pleasing sound.

The fundamental frequency $\frac{\pi a}{L}$ increases as we decrease length L . That is, if we place a finger on the fingerboard and then pluck a string we get a higher note. The constant a is given by

$$a = \sqrt{\frac{T}{\rho}},$$

where T is tension and ρ is the linear density of the string. Tightening the string (turning the tuning peg on a guitar) increases a and hence produces a higher fundamental frequency (a higher note). On the other hand using a heavier string reduces a and produces a lower fundamental frequency (a lower note). A bass guitar has longer thicker strings, while a

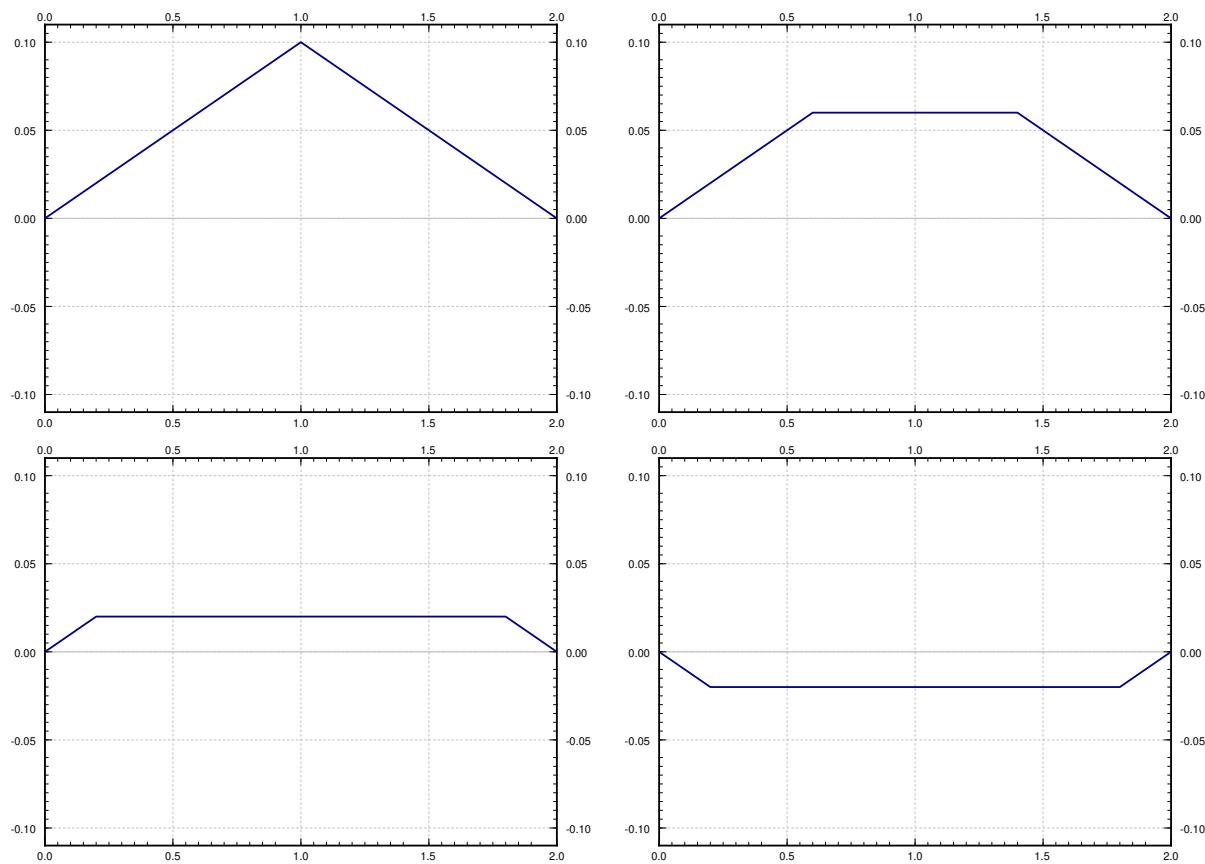


Figure 10.14: Plucked string for $t = 0$, $t = 0.4$, $t = 0.8$, and $t = 1.2$.

ukulele has short strings made of lighter material.

Something rather interesting is the almost symmetry between space and time. In its simplest form we see this symmetry in the solutions

$$\sin\left(\frac{n\pi}{L}x\right) \sin\left(\frac{n\pi a}{L}t\right).$$

Except for the a , time and space are just the same.

In general, the solution for a fixed x is a Fourier series in t , for a fixed t it is a Fourier series in x , and the coefficients are related. If the shape $f(x)$ or the initial velocity have lots of corners, then the sound wave will have lots of corners. That is because the Fourier coefficients of the initial shape decay to zero (as $n \rightarrow \infty$) at the same rate as the Fourier coefficients of the wave in time (for some fixed x). So if you use a sharp object to pick the string, you get a sharper sound with lots of high frequency components, while if you use your thumb, you get a softer sound without so many high overtones. Similarly if you pluck close to the bridge, you are getting a pluck that looks more like the sawtooth, and you get an even sharper sound.

In fact, if you look at the formula for the solution, you see that for any fixed x we get an almost arbitrary Fourier series in t , everything except the constant term. You can essentially

obtain any sound you want by plucking the string in just the right way. Of course we are considering an ideal string of no stiffness and no air resistance. Those variables clearly impact the sound as well.

10.4.3 Exercises

Exercise 10.4.3: Solve

$$\begin{aligned} y_{tt} &= 9y_{xx}, \\ y(0, t) &= y(1, t) = 0, \\ y(x, 0) &= \sin(3\pi x) + \frac{1}{4} \sin(6\pi x) && \text{for } 0 < x < 1, \\ y_t(x, 0) &= 0 && \text{for } 0 < x < 1. \end{aligned}$$

Exercise 10.4.4:* Solve

$$\begin{aligned} y_{tt} &= y_{xx}, \\ y(0, t) &= y(\pi, t) = 0, \\ y(x, 0) &= \sin(x) && \text{for } 0 < x < \pi, \\ y_t(x, 0) &= \sin(x) && \text{for } 0 < x < \pi. \end{aligned}$$

Exercise 10.4.5: Solve

$$\begin{aligned} y_{tt} &= 4y_{xx}, \\ y(0, t) &= y(1, t) = 0, \\ y(x, 0) &= \sin(3\pi x) + \frac{1}{4} \sin(6\pi x) && \text{for } 0 < x < 1, \\ y_t(x, 0) &= \sin(9\pi x) && \text{for } 0 < x < 1. \end{aligned}$$

Exercise 10.4.6:* Solve

$$\begin{aligned} y_{tt} &= 25y_{xx}, \\ y(0, t) &= y(2, t) = 0, \\ y(x, 0) &= 0 && \text{for } 0 < x < 2, \\ y_t(x, 0) &= \sin(\pi t) + 0.1 \sin(2\pi t) && \text{for } 0 < x < 2. \end{aligned}$$

Exercise 10.4.7:* Solve

$$\begin{aligned} y_{tt} &= 2y_{xx}, \\ y(0, t) &= y(\pi, t) = 0, \\ y(x, 0) &= x && \text{for } 0 < x < \pi, \\ y_t(x, 0) &= 0 && \text{for } 0 < x < \pi. \end{aligned}$$

Exercise 10.4.8: Derive the solution for a general plucked string of length L and any constant a (in the equation $y_{tt} = a^2 y_{xx}$), where we raise the string some distance b at the midpoint and let go.

Exercise 10.4.9: Imagine that a stringed musical instrument falls on the floor. Suppose that the length of the string is 1 and $a = 1$. When the musical instrument hits the ground the string was in rest position and hence $y(x, 0) = 0$. However, the string was moving at some velocity at impact ($t = 0$), say $y_t(x, 0) = -1$. Find the solution $y(x, t)$ for the shape of the string at time t .

Exercise 10.4.10:* Let's see what happens when $a = 0$. Find a solution to $y_{tt} = 0$, $y(0, t) = y(\pi, t) = 0$, $y(x, 0) = \sin(2x)$, $y_t(x, 0) = \sin(x)$.

Exercise 10.4.11 (challenging): Suppose that you have a vibrating string and that there is air resistance proportional to the velocity. That is, you have

$$\begin{aligned}y_{tt} &= a^2 y_{xx} - ky_t, \\y(0, t) &= y(1, t) = 0, \\y(x, 0) &= f(x) && \text{for } 0 < x < 1, \\y_t(x, 0) &= 0 && \text{for } 0 < x < 1.\end{aligned}$$

Suppose that $0 < k < 2\pi a$. Derive a series solution to the problem. Any coefficients in the series should be expressed as integrals of $f(x)$.

Exercise 10.4.12: Suppose you touch the guitar string exactly in the middle to ensure another condition $u(L/2, t) = 0$ for all time. Which multiples of the fundamental frequency $\frac{\pi a}{L}$ show up in the solution?

10.5 D'Alembert solution of the wave equation

Attribution: [JL], §4.8.

Learning Objectives

After this section, you will be able to:

- Use D'Alembert solutions to solve the wave equation.

We have solved the wave equation by using Fourier series. But it is often more convenient to use the so-called *d'Alembert solution to the wave equation**. While this solution can be derived using Fourier series as well, it is really an awkward use of those concepts. It is easier and more instructive to derive this solution by making a correct change of variables to get an equation that can be solved by simple integration.

Suppose we wish to solve the wave equation

$$y_{tt} = a^2 y_{xx} \quad (10.10)$$

subject to the side conditions

$$\begin{aligned} y(0, t) &= y(L, t) = 0 && \text{for all } t, \\ y(x, 0) &= f(x) && 0 < x < L, \\ y_t(x, 0) &= g(x) && 0 < x < L. \end{aligned} \quad (10.11)$$

10.5.1 Change of variables

We will transform the equation into a simpler form where it can be solved by simple integration. We change variables to $\xi = x - at$, $\eta = x + at$. The chain rule says:

$$\begin{aligned} \frac{\partial}{\partial x} &= \frac{\partial \xi}{\partial x} \frac{\partial}{\partial \xi} + \frac{\partial \eta}{\partial x} \frac{\partial}{\partial \eta} = \frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}, \\ \frac{\partial}{\partial t} &= \frac{\partial \xi}{\partial t} \frac{\partial}{\partial \xi} + \frac{\partial \eta}{\partial t} \frac{\partial}{\partial \eta} = -a \frac{\partial}{\partial \xi} + a \frac{\partial}{\partial \eta}. \end{aligned}$$

We compute

$$\begin{aligned} y_{xx} &= \frac{\partial^2 y}{\partial x^2} = \left(\frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta} \right) \left(\frac{\partial y}{\partial \xi} + \frac{\partial y}{\partial \eta} \right) = \frac{\partial^2 y}{\partial \xi^2} + 2 \frac{\partial^2 y}{\partial \xi \partial \eta} + \frac{\partial^2 y}{\partial \eta^2}, \\ y_{tt} &= \frac{\partial^2 y}{\partial t^2} = \left(-a \frac{\partial}{\partial \xi} + a \frac{\partial}{\partial \eta} \right) \left(-a \frac{\partial y}{\partial \xi} + a \frac{\partial y}{\partial \eta} \right) = a^2 \frac{\partial^2 y}{\partial \xi^2} - 2a^2 \frac{\partial^2 y}{\partial \xi \partial \eta} + a^2 \frac{\partial^2 y}{\partial \eta^2}. \end{aligned}$$

In the computations above, we used the fact from calculus that $\frac{\partial^2 y}{\partial \xi \partial \eta} = \frac{\partial^2 y}{\partial \eta \partial \xi}$. We plug what we got into the wave equation,

$$0 = a^2 y_{xx} - y_{tt} = 4a^2 \frac{\partial^2 y}{\partial \xi \partial \eta} = 4a^2 y_{\xi \eta}.$$

*Named after the French mathematician Jean le Rond d'Alembert (1717–1783).

Therefore, the wave equation (10.10) transforms into $y_{\xi\eta} = 0$. It is easy to find the general solution to this equation by integrating twice. Keeping ξ constant, we integrate with respect to η first* and notice that the constant of integration depends on ξ ; for each ξ we might get a different constant of integration. We get $y_\xi = C(\xi)$. Next, we integrate with respect to ξ and notice that the constant of integration depends on η . Thus, $y = \int C(\xi) d\xi + B(\eta)$. The solution must, therefore, be of the following form for some functions $A(\xi)$ and $B(\eta)$:

$$y = A(\xi) + B(\eta) = A(x - at) + B(x + at).$$

The solution is a superposition of two functions (waves) traveling at speed a in opposite directions. The coordinates ξ and η are called the *characteristic coordinates*, and a similar technique can be applied to more complicated hyperbolic PDE. And in fact, in § 10.1 it is used to solve first order linear PDE. Basically, to solve the wave equation (or more general hyperbolic equations) we find certain characteristic curves along which the equation is really just an ODE, or a pair of ODEs. In this case these are the curves where ξ and η are constant.

10.5.2 D'Alembert's formula

We know what any solution must look like, but we need to solve for the given side conditions. We will just give the formula and see that it works. First let $F(x)$ denote the odd periodic extension of $f(x)$, and let $G(x)$ denote the odd periodic extension of $g(x)$. Define

$$A(x) = \frac{1}{2}F(x) - \frac{1}{2a} \int_0^x G(s) ds, \quad B(x) = \frac{1}{2}F(x) + \frac{1}{2a} \int_0^x G(s) ds.$$

We claim this $A(x)$ and $B(x)$ give the solution. Explicitly, the solution is $y(x, t) = A(x - at) + B(x + at)$ or in other words:

$$\begin{aligned} y(x, t) &= \frac{1}{2}F(x - at) - \frac{1}{2a} \int_0^{x-at} G(s) ds + \frac{1}{2}F(x + at) + \frac{1}{2a} \int_0^{x+at} G(s) ds \\ &= \frac{F(x - at) + F(x + at)}{2} + \frac{1}{2a} \int_{x-at}^{x+at} G(s) ds. \end{aligned}$$
(10.12)

Let us check that the d'Alembert formula really works.

$$y(x, 0) = \frac{F(x) + F(x)}{2} + \frac{1}{2a} \int_x^x G(s) ds = F(x).$$

So far so good. Assume for simplicity F is differentiable. And we use the first form of (10.12) as it is easier to differentiate. By the fundamental theorem of calculus we have

$$y_t(x, t) = \frac{-a}{2}F'(x - at) + \frac{1}{2}G(x - at) + \frac{a}{2}F'(x + at) + \frac{1}{2}G(x + at).$$

So

$$y_t(x, 0) = \frac{-a}{2}F'(x) + \frac{1}{2}G(x) + \frac{a}{2}F'(x) + \frac{1}{2}G(x) = G(x).$$

*There is nothing special about η , you can integrate with ξ first, if you wish.

Yay! We're smoking now. OK, now the boundary conditions. Note that $F(x)$ and $G(x)$ are odd. So

$$y(0, t) = \frac{F(-at) + F(at)}{2} + \frac{1}{2a} \int_{-at}^{at} G(s) ds = \frac{-F(at) + F(at)}{2} + \frac{1}{2a} \int_{-at}^{at} G(s) ds = 0 + 0 = 0.$$

Now $F(x)$ is odd and $2L$ -periodic, so

$$F(L - at) + F(L + at) = F(-L - at) + F(L + at) = -F(L + at) + F(L + at) = 0.$$

Next, $G(s)$ is odd and $2L$ -periodic, so we change variables $v = s - L$. We then notice that $G(v + L) = G(v - L) = -G(-v + L)$, so $G(v + L)$ is odd as a function of v :

$$\int_{L-at}^{L+at} G(s) ds = \int_{-at}^{at} G(v + L) dv = 0.$$

Hence

$$y(L, t) = \frac{F(L - at) + F(L + at)}{2} + \frac{1}{2a} \int_{L-at}^{L+at} G(s) ds = 0 + 0 = 0.$$

And voilà, it works.

Example 10.5.1: D'Alembert says that the solution is a superposition of two functions (waves) moving in the opposite direction at “speed” a . To get an idea of how it works, let us work out an example. Consider the simpler setup

$$\begin{aligned} y_{tt} &= y_{xx}, \\ y(0, t) &= y(1, t) = 0, \\ y(x, 0) &= f(x), \\ y_t(x, 0) &= 0. \end{aligned}$$

Here $f(x)$ is an impulse of height 1 centered at $x = 0.5$:

$$f(x) = \begin{cases} 0 & \text{if } 0 \leq x < 0.45, \\ 20(x - 0.45) & \text{if } 0.45 \leq x < 0.5, \\ 20(0.55 - x) & \text{if } 0.5 \leq x < 0.55, \\ 0 & \text{if } 0.55 \leq x \leq 1. \end{cases}$$

The graph of this impulse is the top left plot in [Figure 10.15](#) on the next page.

Solution: Let $F(x)$ be the odd periodic extension of $f(x)$. Then (10.12) says that the solution is

$$y(x, t) = \frac{F(x - t) + F(x + t)}{2}.$$

It is not hard to compute specific values of $y(x, t)$. For example, to compute $y(0.1, 0.6)$ we notice $x - t = -0.5$ and $x + t = 0.7$. Now $F(-0.5) = -f(0.5) = -20(0.55 - 0.5) = -1$ and $F(0.7) = f(0.7) = 0$. Hence $y(0.1, 0.6) = \frac{-1+0}{2} = -0.5$. As you can see the d'Alembert solution is much easier to actually compute and to plot than the Fourier series solution. See [Figure 10.15](#) on the facing page for plots of the solution y for several different t . □

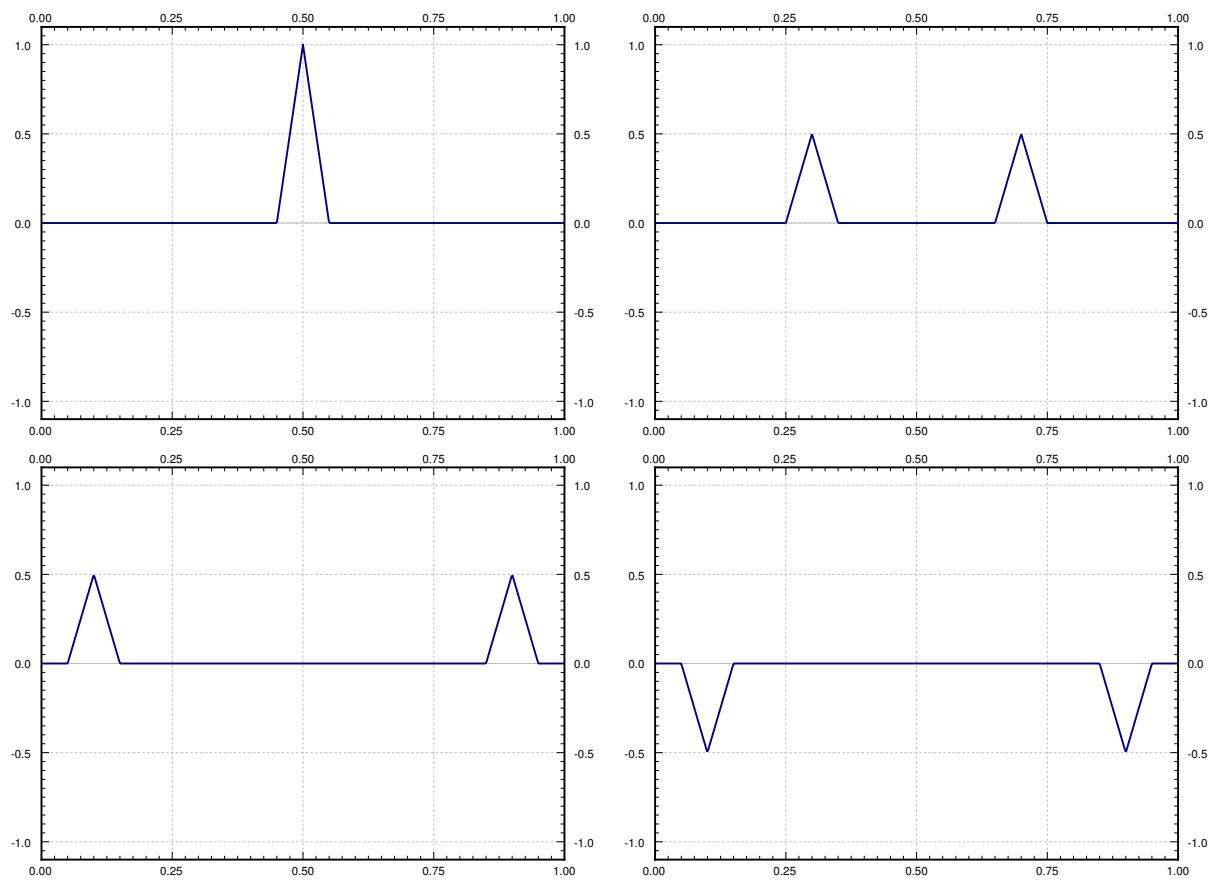


Figure 10.15: Plot of the d'Alembert solution for $t = 0$, $t = 0.2$, $t = 0.4$, and $t = 0.6$.

10.5.3 Another way to solve for the side conditions

It is perhaps easier and more useful to memorize the procedure rather than the formula itself. The important thing to remember is that a solution to the wave equation is a superposition of two waves traveling in opposite directions. That is,

$$y(x, t) = A(x - at) + B(x + at).$$

If you think about it, the exact formulas for A and B are not hard to guess once you realize what kind of side conditions $y(x, t)$ is supposed to satisfy. Let us find the formula again, but slightly differently. Best approach is to do it in stages. When $g(x) = 0$ (and hence $G(x) = 0$) the solution is

$$\frac{F(x - at) + F(x + at)}{2}.$$

On the other hand, when $f(x) = 0$ (and hence $F(x) = 0$), we let

$$H(x) = \int_0^x G(s) ds.$$

The solution in this case is

$$\frac{1}{2a} \int_{x-at}^{x+at} G(s) ds = \frac{-H(x-at) + H(x+at)}{2a}.$$

By superposition we get a solution for the general side conditions (10.11) (when neither $f(x)$ nor $g(x)$ are identically zero).

$$y(x, t) = \frac{F(x-at) + F(x+at)}{2} + \frac{-H(x-at) + H(x+at)}{2a}. \quad (10.13)$$

Do note the minus sign before the H , and the a in the second denominator.

Exercise 10.5.1: Check that the new formula (10.13) satisfies the side conditions (10.11).

Warning: Make sure you use the odd periodic extensions $F(x)$ and $G(x)$, when you have formulas for $f(x)$ and $g(x)$. The thing is, those formulas in general hold only for $0 < x < L$, and are not usually equal to $F(x)$ and $G(x)$ for other x .

10.5.4 Some remarks

Let us remark that the formula $y(x, t) = A(x-at) + B(x+at)$ is the reason why the solution of the wave equation doesn't get "nicer" as time goes on, that is, why in the examples where the initial conditions had corners, the solution also has corners at every time t .

The corners bring us to another interesting remark. Nobody ever notices at first that our example solutions are not even differentiable (they have corners): In Example 10.5.1 above, the solution is not differentiable whenever $x = t + 0.5$ or $x = -t + 0.5$ for example. Really to be able to compute u_{xx} or u_{tt} , you need not one, but two derivatives. Fear not, we could think of a shape that is very nearly $F(x)$ but does have two derivatives by rounding the corners a little bit, and then the solution would be very nearly $\frac{F(x-t)+F(x+t)}{2}$ and nobody would notice the switch.

One final remark is what the d'Alembert solution tells us about what part of the initial conditions influence the solution at a certain point. We can figure this out by "traveling backwards along the characteristics." Let us suppose that the string is very long (perhaps infinite) for simplicity. Since the solution at time t is

$$y(x, t) = \frac{F(x-at) + F(x+at)}{2} + \frac{1}{2a} \int_{x-at}^{x+at} G(s) ds,$$

we notice that we have only used the initial conditions in the interval $[x-at, x+at]$. These two endpoints are called the *wavefronts*, as that is where the wave front is given an initial ($t=0$) disturbance at x . So if $a=1$, an observer sitting at $x=0$ at time $t=1$ has only seen the initial conditions for x in the range $[-1, 1]$ and is blissfully unaware of anything else. This is why for example we do not know that a supernova has occurred in the universe until we see its light, millions of years from the time when it did in fact happen.

10.5.5 Exercises

Exercise 10.5.2: Using the d'Alembert solution solve $y_{tt} = 4y_{xx}$, $0 < x < \pi$, $t > 0$, $y(0, t) = y(\pi, t) = 0$, $y(x, 0) = \sin x$, and $y_t(x, 0) = \sin x$. Hint: Note that $\sin x$ is the odd periodic extension of $y(x, 0)$ and $y_t(x, 0)$.

Exercise 10.5.3:* Using the d'Alembert solution solve $y_{tt} = 9y_{xx}$, $0 < x < 1$, $t > 0$, $y(0, t) = y(1, t) = 0$, $y(x, 0) = \sin(2\pi x)$, and $y_t(x, 0) = \sin(3\pi x)$.

Exercise 10.5.4: Using the d'Alembert solution solve $y_{tt} = 2y_{xx}$, $0 < x < 1$, $t > 0$, $y(0, t) = y(1, t) = 0$, $y(x, 0) = \sin^5(\pi x)$, and $y_t(x, 0) = \sin^3(\pi x)$.

Exercise 10.5.5: Take $y_{tt} = 4y_{xx}$, $0 < x < \pi$, $t > 0$, $y(0, t) = y(\pi, t) = 0$, $y(x, 0) = x(\pi - x)$, and $y_t(x, 0) = 0$.

- a) Solve using the d'Alembert formula. Hint: You can use the sine series for $y(x, 0)$.
- b) Find the solution as a function of x for a fixed $t = 0.5$, $t = 1$, and $t = 2$. Do not use the sine series here.

Exercise 10.5.6:* Take $y_{tt} = 4y_{xx}$, $0 < x < 1$, $t > 0$, $y(0, t) = y(1, t) = 0$, $y(x, 0) = x - x^2$, and $y_t(x, 0) = 0$. Using the d'Alembert solution find the solution at

$$\text{a) } t = 0.1, \quad \text{b) } t = 1/2, \quad \text{c) } t = 1.$$

You may have to split your answer up by cases.

Exercise 10.5.7: Derive the d'Alembert solution for $y_{tt} = a^2 y_{xx}$, $0 < x < \pi$, $t > 0$, $y(0, t) = y(\pi, t) = 0$, $y(x, 0) = f(x)$, and $y_t(x, 0) = 0$, using the Fourier series solution of the wave equation, by applying an appropriate trigonometric identity. Hint: Do it first for a single term of the Fourier series solution, in particular do it when y is $\sin\left(\frac{n\pi}{L}x\right)\sin\left(\frac{n\pi a}{L}t\right)$.

Exercise 10.5.8: The d'Alembert solution still works if there are no boundary conditions and the initial condition is defined on the whole real line. Suppose that $y_{tt} = y_{xx}$ (for all x on the real line and $t \geq 0$), $y(x, 0) = f(x)$, and $y_t(x, 0) = 0$, where

$$f(x) = \begin{cases} 0 & \text{if } x < -1, \\ x + 1 & \text{if } -1 \leq x < 0, \\ -x + 1 & \text{if } 0 \leq x < 1, \\ 0 & \text{if } 1 < x. \end{cases}$$

Solve using the d'Alembert solution. That is, write down a piecewise definition for the solution. Then sketch the solution for $t = 0$, $t = 1/2$, $t = 1$, and $t = 2$.

Exercise 10.5.9:* Take $y_{tt} = 100y_{xx}$, $0 < x < 4$, $t > 0$, $y(0, t) = y(4, t) = 0$, $y(x, 0) = F(x)$, and $y_t(x, 0) = 0$. Suppose that $F(0) = 0$, $F(1) = 2$, $F(2) = 3$, $F(3) = 1$. Using the d'Alembert solution find

$$\text{a) } y(1, 1), \quad \text{b) } y(4, 3), \quad \text{c) } y(3, 9).$$

10.6 Steady state temperature and the Laplacian

Attribution: [JL], §4.9.

Learning Objectives

After this section, you will be able to:

- Relate the heat equation independent of time to the Laplace equation and
- Use separation of variables to solve the Laplace equation on rectangular regions.

Consider an insulated wire, a plate, or a 3-dimensional object. We apply certain fixed temperatures on the ends of the wire, the edges of the plate, or on all sides of the 3-dimensional object. We wish to find out what is the *steady state temperature* distribution. That is, we wish to know what will be the temperature after long enough period of time.

We are really looking for a solution to the heat equation that is not dependent on time. Let us first solve the problem in one space variable. We are looking for a function u that satisfies

$$u_t = ku_{xx},$$

but such that $u_t = 0$ for all x and t . Hence, we are looking for a function of x alone that satisfies $u_{xx} = 0$. It is easy to solve this equation by integration and we see that $u = Ax + B$ for some constants A and B .

Consider an insulated wire where we apply constant temperature T_1 at one end (say where $x = 0$) and T_2 on the other end (at $x = L$ where L is the length of the wire). Our steady state solution is

$$u(x) = \frac{T_2 - T_1}{L}x + T_1.$$

This solution agrees with our common sense intuition with how the heat should be distributed in the wire. So in one dimension, the steady state solutions are basically just straight lines.

Things are more complicated in two or more space dimensions. Let us restrict to two space dimensions for simplicity. The heat equation in two space variables is

$$u_t = k(u_{xx} + u_{yy}), \quad (10.14)$$

or more commonly written as $u_t = k\Delta u$ or $u_t = k\nabla^2 u$. Here the Δ and ∇^2 symbols mean $\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$. We will use Δ from now on. The reason for using such a notation is that you can define Δ to be the right thing for any number of space dimensions and then the heat equation is always $u_t = k\Delta u$. The operator Δ is called the *Laplacian*.

OK, now that we have notation out of the way, let us see what does an equation for the steady state solution look like. We are looking for a solution to (10.14) that does not depend on t , or in other words $u_t = 0$. Hence we are looking for a function $u(x, y)$ such that

$$\Delta u = u_{xx} + u_{yy} = 0.$$

This equation is called the *Laplace equation**, and is an example of an elliptic equation. Solutions to the Laplace equation are called *harmonic functions* and have many nice properties

*Named after the French mathematician Pierre-Simon, marquis de Laplace (1749–1827).

and applications far beyond the steady state heat problem. One of these main applications is in electrostatics, as the electric potential V in a region also solves the Laplace equation

$$\Delta V = 0.$$

Harmonic functions in two variables are no longer just linear (plane graphs). For example, you can check that the functions $x^2 - y^2$ and xy are harmonic. However, note that if u_{xx} is positive, u is concave up in the x direction, then u_{yy} must be negative and u must be concave down in the y direction. A harmonic function can never have any “hilltop” or “valley” on the graph. This observation is consistent with our intuitive idea of steady state heat distribution; the hottest or coldest spot will not be inside.

Commonly the Laplace equation is part of a so-called *Dirichlet problem*^{*}. That is, we have a region in the xy -plane and we specify certain values along the boundaries of the region. We then try to find a solution u to the Laplace equation defined on this region such that u agrees with the values we specified on the boundary.

In this section we consider a rectangular region. For simplicity we specify boundary values to be zero at 3 of the four edges and only specify an arbitrary function at one edge. As we still have the principle of superposition, we can use this simpler solution to derive the general solution for arbitrary boundary values by solving 4 different problems, one for each edge, and adding those solutions together. This setup is left as an exercise.

We wish to solve the following problem. Let h and w be the height and width of our rectangle, with one corner at the origin and lying in the first quadrant.

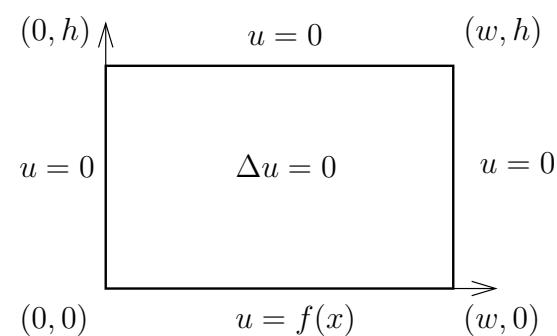
$$\Delta u = 0, \quad (10.15)$$

$$u(0, y) = 0 \quad \text{for } 0 < y < h, \quad (10.16)$$

$$u(x, h) = 0 \quad \text{for } 0 < x < w, \quad (10.17)$$

$$u(w, y) = 0 \quad \text{for } 0 < y < h, \quad (10.18)$$

$$u(x, 0) = f(x) \quad \text{for } 0 < x < w. \quad (10.19)$$



The method we apply is separation of variables. Again, we will come up with enough building-block solutions satisfying all the homogeneous boundary conditions (all conditions except (10.19)). We notice that superposition still works for the equation and all the homogeneous conditions. Therefore, we can use the Fourier series for $f(x)$ to solve the problem as before.

We try $u(x, y) = X(x)Y(y)$. We plug u into the equation to get

$$X''Y + XY'' = 0.$$

We put the X s on one side and the Y s on the other to get

$$-\frac{X''}{X} = \frac{Y''}{Y}.$$

*Named after the German mathematician Johann Peter Gustav Lejeune Dirichlet (1805–1859).

The left-hand side only depends on x and the right-hand side only depends on y . Therefore, there is some constant λ such that $\lambda = \frac{-X''}{X} = \frac{Y''}{Y}$. And we get two equations

$$\begin{aligned} X'' + \lambda X &= 0, \\ Y'' - \lambda Y &= 0. \end{aligned}$$

Furthermore, the homogeneous boundary conditions imply that $X(0) = X(w) = 0$ and $Y(h) = 0$. Taking the equation for X we have already seen that we have a nontrivial solution if and only if $\lambda = \lambda_n = \frac{n^2\pi^2}{w^2}$ and the solution is a multiple of

$$X_n(x) = \sin\left(\frac{n\pi}{w}x\right).$$

For these given λ_n , the general solution for Y (one for each n) is

$$Y_n(y) = A_n \cosh\left(\frac{n\pi}{w}y\right) + B_n \sinh\left(\frac{n\pi}{w}y\right). \quad (10.20)$$

We only have one condition on Y_n and hence we can pick one of A_n or B_n to be something convenient. It will be useful to have $Y_n(0) = 1$, so we let $A_n = 1$. Setting $Y_n(h) = 0$ and solving for B_n we get that

$$B_n = \frac{-\cosh\left(\frac{n\pi h}{w}\right)}{\sinh\left(\frac{n\pi h}{w}\right)}.$$

After we plug the A_n and B_n we into (10.20) and simplify by using the identity $\sinh(\alpha - \beta) = \sinh(\alpha)\cosh(\beta) - \cosh(\alpha)\sinh(\beta)$, we find

$$Y_n(y) = \frac{\sinh\left(\frac{n\pi(h-y)}{w}\right)}{\sinh\left(\frac{n\pi h}{w}\right)}.$$

We define $u_n(x, y) = X_n(x)Y_n(y)$. And note that u_n satisfies (10.15)–(10.18).

Observe that

$$u_n(x, 0) = X_n(x)Y_n(0) = \sin\left(\frac{n\pi}{w}x\right).$$

Suppose

$$f(x) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi}{w}x\right).$$

Then we get a solution of (10.15)–(10.19) of the following form.

$$u(x, y) = \sum_{n=1}^{\infty} b_n u_n(x, y) = \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi}{w}x\right) \left(\frac{\sinh\left(\frac{n\pi(h-y)}{w}\right)}{\sinh\left(\frac{n\pi h}{w}\right)} \right).$$

As u_n satisfies (10.15)–(10.18) and any linear combination (finite or infinite) of u_n also satisfies (10.15)–(10.18), then u satisfies (10.15)–(10.18). By plugging in $y = 0$, we see u satisfies (10.19) as well.

Example 10.6.1: Take $w = h = \pi$ and let $f(x) = \pi$. Let us compute the sine series for the function π (same as the series for the square wave). For $0 < x < \pi$, we have

$$f(x) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{4}{n} \sin(nx).$$

Therefore the solution $u(x, y)$, see Figure 10.16, to the corresponding Dirichlet problem is given as

$$u(x, y) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{4}{n} \sin(nx) \left(\frac{\sinh(n(\pi - y))}{\sinh(n\pi)} \right).$$

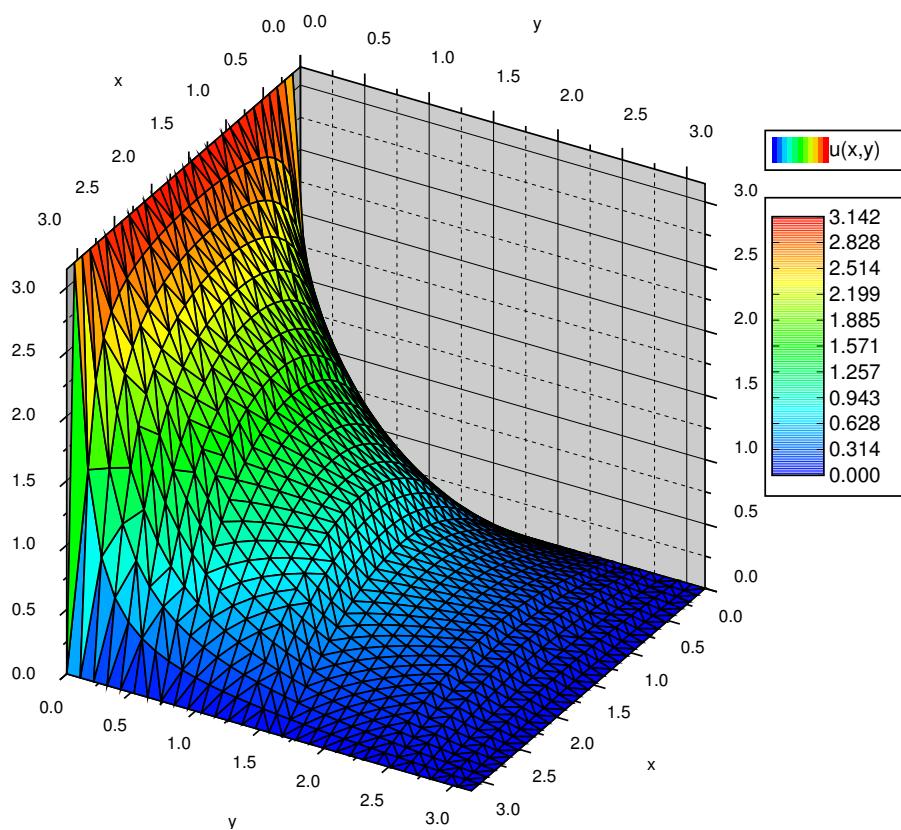


Figure 10.16: Steady state temperature of a square plate, three sides held at zero and one side held at π .

This scenario corresponds to the steady state temperature on a square plate of width π with 3 sides held at 0 degrees and one side held at π degrees. If we have arbitrary initial data on all sides, then we solve four problems, each using one piece of nonhomogeneous data. Then we use the principle of superposition to add up all four solutions to have a solution to the original problem.

A different way to visualize solutions of the Laplace equation is to take a wire and bend it so that it corresponds to the graph of the temperature above the boundary of your region. Cut a rubber sheet in the shape of your region—a square in our case—and stretch it fixing the edges of the sheet to the wire. The rubber sheet is a good approximation of the graph of the solution to the Laplace equation with the given boundary data.

10.6.1 Exercises

Exercise 10.6.1: Let R be the region described by $0 < x < \pi$ and $0 < y < \pi$. Solve the problem

$$\Delta u = 0, \quad u(x, 0) = \sin x, \quad u(x, \pi) = 0, \quad u(0, y) = 0, \quad u(\pi, y) = 0.$$

Exercise 10.6.2: Let R be the region described by $0 < x < 1$ and $0 < y < 1$. Solve the problem

$$\begin{aligned} u_{xx} + u_{yy} &= 0, \\ u(x, 0) &= \sin(\pi x) - \sin(2\pi x), \quad u(x, 1) = 0, \\ u(0, y) &= 0, \quad u(1, y) = 0. \end{aligned}$$

Exercise 10.6.3: Let R be the region described by $0 < x < 1$ and $0 < y < 1$. Solve the problem

$$\begin{aligned} u_{xx} + u_{yy} &= 0, \\ u(x, 0) &= u(x, 1) = u(0, y) = u(1, y) = C. \end{aligned}$$

for some constant C . Hint: Guess, then check your intuition.

Exercise 10.6.4: Let R be the region described by $0 < x < \pi$ and $0 < y < \pi$. Solve

$$\Delta u = 0, \quad u(x, 0) = 0, \quad u(x, \pi) = \pi, \quad u(0, y) = y, \quad u(\pi, y) = y.$$

Hint: Try a solution of the form $u(x, y) = X(x) + Y(y)$ (different separation of variables).

Exercise 10.6.5: Use the solution of [Exercise 10.6.4](#) to solve

$$\Delta u = 0, \quad u(x, 0) = \sin x, \quad u(x, \pi) = \pi, \quad u(0, y) = y, \quad u(\pi, y) = y.$$

Hint: Use superposition.

Exercise 10.6.6: Let R be the region described by $0 < x < w$ and $0 < y < h$. Solve the problem

$$\begin{aligned} u_{xx} + u_{yy} &= 0, \\ u(x, 0) &= 0, \quad u(x, h) = f(x), \\ u(0, y) &= 0, \quad u(w, y) = 0. \end{aligned}$$

The solution should be in series form using the Fourier series coefficients of $f(x)$.

Exercise 10.6.7:* Let R be the region described by $0 < x < 1$ and $0 < y < 1$. Solve the problem

$$\Delta u = 0, \quad u(x, 0) = \sum_{n=1}^{\infty} \frac{1}{n^2} \sin(n\pi x), \quad u(x, 1) = 0, \quad u(0, y) = 0, \quad u(1, y) = 0.$$

Exercise 10.6.8: Let R be the region described by $0 < x < w$ and $0 < y < h$. Solve the problem

$$\begin{aligned} u_{xx} + u_{yy} &= 0, \\ u(x, 0) &= 0, \quad u(x, h) = 0, \\ u(0, y) &= f(y), \quad u(w, y) = 0. \end{aligned}$$

The solution should be in series form using the Fourier series coefficients of $f(y)$.

Exercise 10.6.9: Let R be the region described by $0 < x < w$ and $0 < y < h$. Solve the problem

$$\begin{aligned} u_{xx} + u_{yy} &= 0, \\ u(x, 0) &= 0, \quad u(x, h) = 0, \\ u(0, y) &= 0, \quad u(w, y) = f(y). \end{aligned}$$

The solution should be in series form using the Fourier series coefficients of $f(y)$.

Exercise 10.6.10: Let R be the region described by $0 < x < 1$ and $0 < y < 1$. Solve the problem

$$\begin{aligned} u_{xx} + u_{yy} &= 0, \\ u(x, 0) &= \sin(9\pi x), \quad u(x, 1) = \sin(2\pi x), \\ u(0, y) &= 0, \quad u(1, y) = 0. \end{aligned}$$

Hint: Use superposition.

Exercise 10.6.11:* Let R be the region described by $0 < x < 1$ and $0 < y < 2$. Solve the problem

$$\Delta u = 0, \quad u(x, 0) = 0.1 \sin(\pi x), \quad u(x, 2) = 0, \quad u(0, y) = 0, \quad u(1, y) = 0.$$

Exercise 10.6.12: Let R be the region described by $0 < x < 1$ and $0 < y < 1$. Solve the problem

$$\begin{aligned} u_{xx} + u_{yy} &= 0, \\ u(x, 0) &= \sin(\pi x), \quad u(x, 1) = \sin(\pi x), \\ u(0, y) &= \sin(\pi y), \quad u(1, y) = \sin(\pi y). \end{aligned}$$

Hint: Use superposition.

Exercise 10.6.13 (challenging): Using only your intuition find $u(1/2, 1/2)$, for the problem $\Delta u = 0$, where $u(0, y) = u(1, y) = 100$ for $0 < y < 1$, and $u(x, 0) = u(x, 1) = 0$ for $0 < x < 1$. Explain.

10.7 Dirichlet problem in the circle and the Poisson kernel

Attribution: [JL], §4.10.

Learning Objectives

After this section, you will be able to:

- Solve the Laplace equation in a circle,
- Use series solutions in polar coordinates to solve the Laplace equation, and
- Use the Poisson kernel to solve boundary value problems for the Laplacian in a circle.

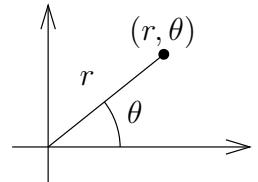
10.7.1 Laplace in polar coordinates

A more natural setting for the Laplace equation $\Delta u = 0$ is a circle rather than a rectangle. On the other hand, what makes the problem somewhat more difficult is that we need polar coordinates.

Recall that the polar coordinates for the (x, y) -plane are (r, θ) :

$$x = r \cos \theta, \quad y = r \sin \theta,$$

where $r \geq 0$ and $-\pi < \theta \leq \pi$. So the point (x, y) is distance r from the origin at an angle θ from the positive x -axis.



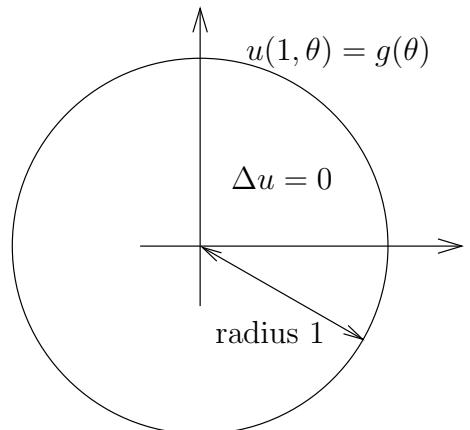
Now that we know our coordinates, let us give the problem we wish to solve. We have a circular region of radius 1, and we are interested in the Dirichlet problem for the Laplace equation for this region. Let $u(r, \theta)$ denote the temperature at the point (r, θ) in polar coordinates.

We have the problem:

$$\begin{aligned} \Delta u &= 0, & \text{for } r < 1, \\ u(1, \theta) &= g(\theta), & \text{for } -\pi < \theta \leq \pi. \end{aligned} \tag{10.21}$$

The first issue we face is that we do not know the Laplacian in polar coordinates. Normally we would find u_{xx} and u_{yy} in terms of the derivatives in r and θ . We would need to solve for r and θ in terms of x and y . In this case it is more convenient to work in reverse. We compute derivatives in r and θ in terms of derivatives in x and y and then we solve. The computations are easier this way. First

$$\begin{aligned} x_r &= \cos \theta, & x_\theta &= -r \sin \theta, \\ y_r &= \sin \theta, & y_\theta &= r \cos \theta. \end{aligned}$$



Next by chain rule we obtain

$$\begin{aligned} u_r &= u_x x_r + u_y y_r = \cos(\theta)u_x + \sin(\theta)u_y, \\ u_{rr} &= \cos(\theta)(u_{xx}x_r + u_{xy}y_r) + \sin(\theta)(u_{yx}x_r + u_{yy}y_r) \\ &= \cos^2(\theta)u_{xx} + 2\cos(\theta)\sin(\theta)u_{xy} + \sin^2(\theta)u_{yy}. \end{aligned}$$

Similarly for the θ derivative. Note that we have to use the product rule for the second derivative.

$$\begin{aligned} u_\theta &= u_x x_\theta + u_y y_\theta = -r\sin(\theta)u_x + r\cos(\theta)u_y, \\ u_{\theta\theta} &= -r\cos(\theta)u_x - r\sin(\theta)(u_{xx}x_\theta + u_{xy}y_\theta) - r\sin(\theta)u_y + r\cos(\theta)(u_{yx}x_\theta + u_{yy}y_\theta) \\ &= -r\cos(\theta)u_x - r\sin(\theta)u_y + r^2\sin^2(\theta)u_{xx} - r^22\sin(\theta)\cos(\theta)u_{xy} + r^2\cos^2(\theta)u_{yy}. \end{aligned}$$

Let us now try to solve for $u_{xx} + u_{yy}$. We start with $\frac{1}{r^2}u_{\theta\theta}$ to get rid of those pesky r^2 . If we add u_{rr} and use the fact that $\cos^2(\theta) + \sin^2(\theta) = 1$, we get

$$\frac{1}{r^2}u_{\theta\theta} + u_{rr} = u_{xx} + u_{yy} - \frac{1}{r}\cos(\theta)u_x - \frac{1}{r}\sin(\theta)u_y.$$

We're not quite there yet, but all we are lacking is $\frac{1}{r}u_r$. Adding it we obtain the *Laplacian in polar coordinates*:

$$\boxed{\Delta u = u_{xx} + u_{yy} = \frac{1}{r^2}u_{\theta\theta} + \frac{1}{r}u_r + u_{rr}.}$$

Notice that the Laplacian in polar coordinates no longer has constant coefficients.

10.7.2 Series solution

Let us separate variables as usual. That is let us try $u(r, \theta) = R(r)\Theta(\theta)$. Then

$$0 = \Delta u = \frac{1}{r^2}R\Theta'' + \frac{1}{r}R'\Theta + R''\Theta.$$

Let us put R on one side and Θ on the other and conclude that both sides must be constant.

$$\begin{aligned} \frac{1}{r^2}R\Theta'' &= -\left(\frac{1}{r}R' + R''\right)\Theta \\ \frac{\Theta''}{\Theta} &= -\frac{rR' + r^2R''}{R} = -\lambda \end{aligned}$$

We get two equations:

$$\begin{aligned} \Theta'' + \lambda\Theta &= 0, \\ r^2R'' + rR' - \lambda R &= 0. \end{aligned}$$

Let us first focus on Θ . We know that $u(r, \theta)$ ought to be 2π -periodic in θ , that is, $u(r, \theta) = u(r, \theta + 2\pi)$. Therefore, the solution to $\Theta'' + \lambda\Theta = 0$ must be 2π -periodic. We have seen such a problem in [Example 9.1.5](#). We conclude that $\lambda = n^2$ for a nonnegative integer

$n = 0, 1, 2, 3, \dots$. The equation becomes $\Theta'' + n^2\Theta = 0$. When $n = 0$ the equation is just $\Theta'' = 0$, so we have the general solution $A\theta + B$. As Θ is periodic, $A = 0$. For convenience we write this solution as

$$\Theta_0 = \frac{a_0}{2}$$

for some constant a_0 . For positive n , the solution to $\Theta'' + n^2\Theta = 0$ is

$$\Theta_n = a_n \cos(n\theta) + b_n \sin(n\theta),$$

for some constants a_n and b_n .

Next, we consider the equation for R ,

$$r^2R'' + rR' - n^2R = 0.$$

This equation appeared in exercises before—we solved it in [Exercise 2.1.26](#) and [Exercise 2.1.27](#) on page 110. The idea is to try a solution r^s and if that does not give us two solutions, also try a solution of the form $r^s \ln r$. Let us name the solution for R_n . When $n = 0$ we obtain

$$R_0 = Ar^0 + Br^0 \ln r = A + B \ln r,$$

and if $n > 0$, we get

$$R_n = Ar^n + Br^{-n}.$$

The function $u(r, \theta)$ must be finite at the origin, that is, when $r = 0$. So $B = 0$ in both cases. Set $A = 1$ in both cases as well; the constants in Θ_n will pick up the slack so nothing is lost. Let

$$R_0 = 1, \quad \text{and} \quad R_n = r^n.$$

Hence our building block solutions are

$$u_0(r, \theta) = \frac{a_0}{2}, \quad u_n(r, \theta) = a_n r^n \cos(n\theta) + b_n r^n \sin(n\theta).$$

Putting everything together our solution is:

$$u(r, \theta) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n r^n \cos(n\theta) + b_n r^n \sin(n\theta).$$

We look at the boundary condition in [\(10.21\)](#),

$$g(\theta) = u(1, \theta) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\theta) + b_n \sin(n\theta).$$

Therefore, to solve [\(10.21\)](#) we expand $g(\theta)$, which is a 2π -periodic function, as a Fourier series, and then multiply the n^{th} term by r^n . To find the a_n and the b_n we compute

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} g(\theta) \cos(n\theta) d\theta, \quad \text{and} \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} g(\theta) \sin(n\theta) d\theta.$$

Example 10.7.1: Suppose we wish to solve

$$\begin{aligned}\Delta u = 0, \quad & 0 \leq r < 1, \quad -\pi < \theta \leq \pi, \\ u(1, \theta) = \cos(10\theta), \quad & -\pi < \theta \leq \pi.\end{aligned}$$

The solution is

$$u(r, \theta) = r^{10} \cos(10\theta).$$

See the plot in [Figure 10.17](#). The thing to notice in this example is that the effect of a high frequency is mostly felt at the boundary. In the middle of the disc, the solution is very close to zero. That is because r^{10} is rather small when r is close to 0.

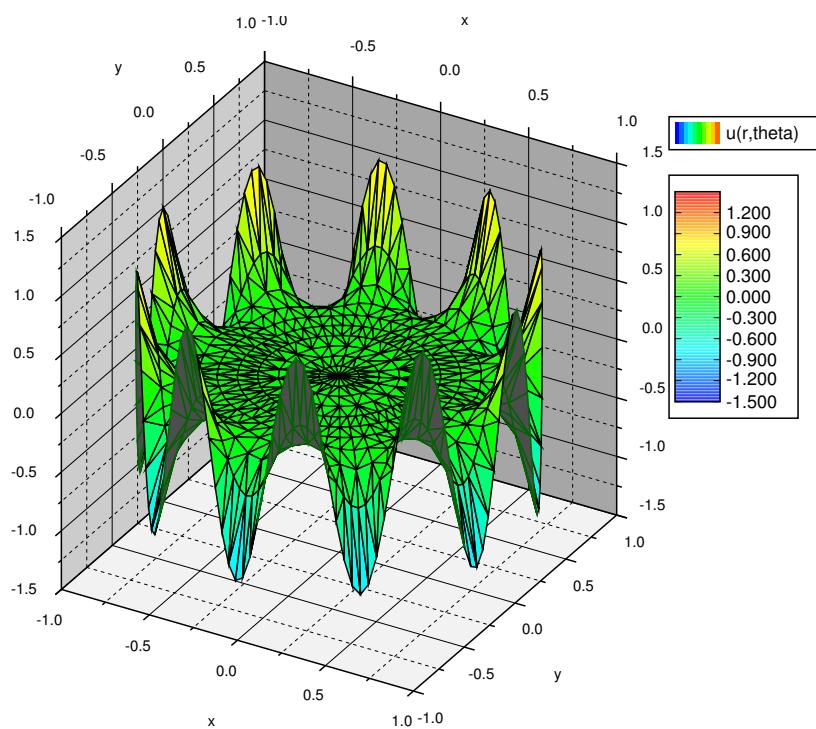


Figure 10.17: The solution of the Dirichlet problem in the disc with $\cos(10\theta)$ as boundary data.

Example 10.7.2: Let us solve a more difficult problem. Consider a long rod with circular cross section of radius 1. Suppose we wish to solve the steady state heat problem in the rod. If the rod is long enough, we simply need to solve the Laplace equation in two dimensions. Let us put the center of the rod at the origin and we have exactly the region we are currently studying—a circle of radius 1. For the boundary conditions, suppose in Cartesian coordinates x and y , the temperature on the boundary is 0 when $y < 0$, and it is $2y$ when $y > 0$.

Let us set the problem up. As $y = r \sin(\theta)$, then on the circle of radius 1, that is, where

$r = 1$, we have $2y = 2\sin(\theta)$. So

$$\Delta u = 0, \quad 0 \leq r < 1, \quad -\pi < \theta \leq \pi,$$

$$u(1, \theta) = \begin{cases} 2\sin(\theta) & \text{if } 0 \leq \theta \leq \pi, \\ 0 & \text{if } -\pi < \theta < 0. \end{cases}$$

We must now compute the Fourier series for the boundary condition. By now the reader has plentiful experience in computing Fourier series and so we simply state that

$$u(1, \theta) = \frac{2}{\pi} + \sin(\theta) + \sum_{n=1}^{\infty} \frac{-4}{\pi(4n^2 - 1)} \cos(2n\theta).$$

Exercise 10.7.1: Compute the series for $u(1, \theta)$ and verify that it really is what we have just claimed. Hint: Be careful, make sure not to divide by zero.

We now simply write the solution (see Figure 10.18) by multiplying by r^n in the right places.

$$u(r, \theta) = \frac{2}{\pi} + r \sin(\theta) + \sum_{n=1}^{\infty} \frac{-4r^{2n}}{\pi(4n^2 - 1)} \cos(2n\theta).$$

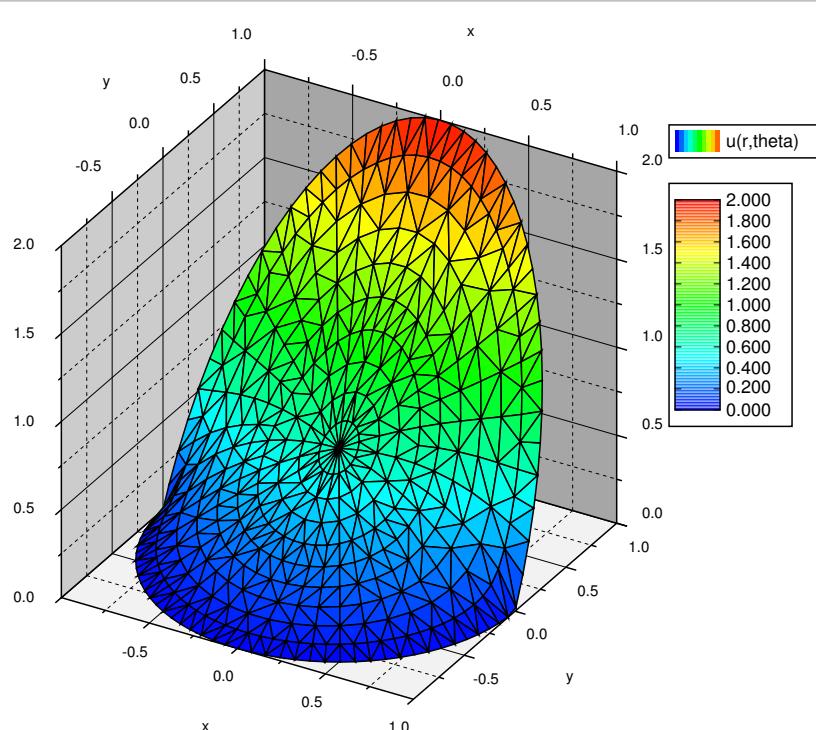


Figure 10.18: The solution of the Dirichlet problem with boundary data 0 for $y < 0$ and $2y$ for $y > 0$.

10.7.3 Poisson kernel

There is another way to solve the Dirichlet problem with the help of an integral kernel. That is, we will find a function $P(r, \theta, \alpha)$ called the *Poisson kernel** such that

$$u(r, \theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(r, \theta, \alpha) g(\alpha) d\alpha.$$

While the integral will generally not be solvable analytically, it can be evaluated numerically. In fact, unless the boundary data is given as a Fourier series already, it may be much easier to numerically evaluate this formula as there is only one integral to evaluate.

The formula also has theoretical applications. For instance, as $P(r, \theta, \alpha)$ will have infinitely many derivatives, then via differentiating under the integral we find that the solution $u(r, \theta)$ has infinitely many derivatives, at least when inside the circle, $r < 1$. By “having infinitely many derivatives,” what you should think of is that $u(r, \theta)$ has “no corners” and all of its partial derivatives of all orders exist and also have “no corners.”

We will compute the formula for $P(r, \theta, \alpha)$ from the series solution, and this idea can be applied anytime you have a convenient series solution where the coefficients are obtained via integration. Hence you can apply this reasoning to obtain such integral kernels for other equations, such as the heat equation. The computation is long and tedious, but not overly difficult. Since the ideas are often applied in similar contexts, it is good to understand how this computation works.

What we do is start with the series solution and replace the coefficients with the integrals that compute them. Then we try to write everything as a single integral. We must use a different dummy variable for the integration and hence we use α instead of θ .

$$\begin{aligned} u(r, \theta) &= \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n r^n \cos(n\theta) + b_n r^n \sin(n\theta) \\ &= \underbrace{\left(\frac{1}{2\pi} \int_{-\pi}^{\pi} g(\alpha) d\alpha \right)}_{a_0/2} + \sum_{n=1}^{\infty} \underbrace{\left(\frac{1}{\pi} \int_{-\pi}^{\pi} g(\alpha) \cos(n\alpha) d\alpha \right)}_{a_n} r^n \cos(n\theta) + \\ &\quad + \underbrace{\left(\frac{1}{\pi} \int_{-\pi}^{\pi} g(\alpha) \sin(n\alpha) d\alpha \right)}_{b_n} r^n \sin(n\theta) \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(g(\alpha) + 2 \sum_{n=1}^{\infty} g(\alpha) \cos(n\alpha) r^n \cos(n\theta) + g(\alpha) \sin(n\alpha) r^n \sin(n\theta) \right) d\alpha \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \underbrace{\left(1 + 2 \sum_{n=1}^{\infty} r^n (\cos(n\alpha) \cos(n\theta) + \sin(n\alpha) \sin(n\theta)) \right)}_{P(r, \theta, \alpha)} g(\alpha) d\alpha \end{aligned}$$

OK, so we have what we wanted, the expression in the parentheses is the Poisson kernel, $P(r, \theta, \alpha)$. However, we can do a lot better. It is still given as a series, and we would really

*Named for the French mathematician Siméon Denis Poisson (1781–1840).

like to have a nice simple expression for it. We must work a little harder. The trick is to rewrite everything in terms of complex exponentials. Let us work just on the kernel.

$$\begin{aligned}
 P(r, \theta, \alpha) &= 1 + 2 \sum_{n=1}^{\infty} r^n (\cos(n\alpha) \cos(n\theta) + \sin(n\alpha) \sin(n\theta)) \\
 &= 1 + 2 \sum_{n=1}^{\infty} r^n \cos(n(\theta - \alpha)) \\
 &= 1 + \sum_{n=1}^{\infty} r^n (e^{in(\theta-\alpha)} + e^{-in(\theta-\alpha)}) \\
 &= 1 + \sum_{n=1}^{\infty} (re^{i(\theta-\alpha)})^n + \sum_{n=1}^{\infty} (re^{-i(\theta-\alpha)})^n.
 \end{aligned}$$

In the expression above, we recognize the *geometric series*. Recall from calculus that if z is a complex number where $|z| < 1$, then

$$\sum_{n=1}^{\infty} z^n = \frac{z}{1-z}.$$

Note that n starts at 1 and that is why we have the z in the numerator. It is the standard geometric series multiplied by z . We can use $z = re^{i(\theta-\alpha)}$, as lo and behold $|re^{i(\theta-\alpha)}| = r < 1$. Let us continue with the computation.

$$\begin{aligned}
 P(r, \theta, \alpha) &= 1 + \sum_{n=1}^{\infty} (re^{i(\theta-\alpha)})^n + \sum_{n=1}^{\infty} (re^{-i(\theta-\alpha)})^n \\
 &= 1 + \frac{re^{i(\theta-\alpha)}}{1 - re^{i(\theta-\alpha)}} + \frac{re^{-i(\theta-\alpha)}}{1 - re^{-i(\theta-\alpha)}} \\
 &= \frac{(1 - re^{i(\theta-\alpha)})(1 - re^{-i(\theta-\alpha)}) + (1 - re^{-i(\theta-\alpha)})re^{i(\theta-\alpha)} + (1 - re^{i(\theta-\alpha)})re^{-i(\theta-\alpha)}}{(1 - re^{i(\theta-\alpha)})(1 - re^{-i(\theta-\alpha)})} \\
 &= \frac{1 - r^2}{1 - re^{i(\theta-\alpha)} - re^{-i(\theta-\alpha)} + r^2} \\
 &= \frac{1 - r^2}{1 - 2r \cos(\theta - \alpha) + r^2}.
 \end{aligned}$$

That's a formula we can live with. The solution to the Dirichlet problem using the Poisson kernel is

$$u(r, \theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1 - r^2}{1 - 2r \cos(\theta - \alpha) + r^2} g(\alpha) d\alpha.$$

Sometimes the formula for the Poisson kernel is given together with the constant $\frac{1}{2\pi}$, in which case we should of course not leave it in front of the integral. Also, often the limits of the integral are given as 0 to 2π ; everything inside is 2π -periodic in α , so this does not change the integral.

Let us not leave the Poisson kernel without explaining its geometric meaning. Let s be the distance from (r, θ) to $(1, \alpha)$. You may recall from calculus that this distance s in polar coordinates is given precisely by the square root of $1 - 2r \cos(\theta - \alpha) + r^2$. That is, the Poisson kernel is really the formula

$$\frac{1 - r^2}{s^2}.$$

One final note we make about the formula is that it is really a weighted average of the boundary values. First let us look at what happens at the origin, that is when $r = 0$.

$$\begin{aligned} u(0, 0) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1 - 0^2}{1 - 2(0) \cos(\theta - \alpha) + 0^2} g(\alpha) d\alpha \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\alpha) d\alpha. \end{aligned}$$

So $u(0, 0)$ is precisely the average value of $g(\theta)$ and therefore the average value of u on the boundary. This is a general feature of harmonic functions, the value at some point p is equal to the average of the values on a circle centered at p .

What the formula says is that the value of the solution at any point in the circle is a weighted average of the boundary data $g(\theta)$. The kernel is bigger when $(1, \alpha)$ is closer to (r, θ) . Therefore when computing $u(r, \theta)$ we give more weight to the values $g(\alpha)$ when $(1, \alpha)$ is closer to (r, θ) and less weight to the values $g(\alpha)$ when $(1, \alpha)$ far from (r, θ) .

10.7.4 Exercises

Exercise 10.7.2: Using series solve $\Delta u = 0$, $u(1, \theta) = |\theta|$, for $-\pi < \theta \leq \pi$.

Exercise 10.7.3:* Using series solve $\Delta u = 0$, $u(1, \theta) = 1 + \sum_{n=1}^{\infty} \frac{1}{n^2} \sin(n\theta)$.

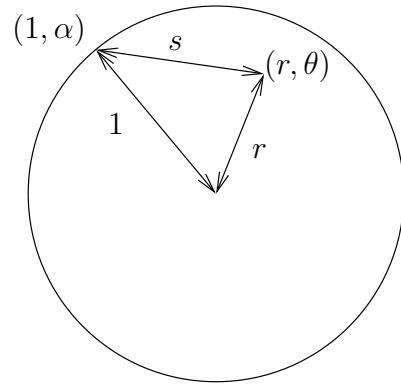
Exercise 10.7.4:* Using the series solution find the solution to $\Delta u = 0$, $u(1, \theta) = 1 - \cos(\theta)$. Express the solution in Cartesian coordinates (that is, using x and y).

Exercise 10.7.5: Using series solve $\Delta u = 0$, $u(1, \theta) = g(\theta)$ for the following data. Hint: trig identities.

a) $g(\theta) = 1/2 + 3 \sin(\theta) + \cos(3\theta)$ b) $g(\theta) = 3 \cos(3\theta) + 3 \sin(3\theta) + \sin(9\theta)$

c) $g(\theta) = 2 \cos(\theta + 1)$ d) $g(\theta) = \sin^2(\theta)$

Exercise 10.7.6: Using the Poisson kernel, give the solution to $\Delta u = 0$, where $u(1, \theta)$ is zero for θ outside the interval $[-\pi/4, \pi/4]$ and $u(1, \theta)$ is 1 for θ on the interval $[-\pi/4, \pi/4]$.



Exercise 10.7.7:

- a) Draw a graph for the Poisson kernel as a function of α when $r = 1/2$ and $\theta = 0$.
- b) Describe what happens to the graph when you make r bigger (as it approaches 1).
- c) Knowing that the solution $u(r, \theta)$ is the weighted average of $g(\theta)$ with Poisson kernel as the weight, explain what your answer to part b) means.

Exercise 10.7.8: Let $g(\theta)$ be the function $xy = \cos \theta \sin \theta$ on the boundary. Use the series solution to find a solution to the Dirichlet problem $\Delta u = 0$, $u(1, \theta) = g(\theta)$. Now convert the solution to Cartesian coordinates x and y . Is this solution surprising? Hint: use your trig identities.

Exercise 10.7.9:*

- a) Try and guess a solution to $\Delta u = -1$, $u(1, \theta) = 0$. Hint: try a solution that only depends on r . Also first, don't worry about the boundary condition.
- b) Now solve $\Delta u = -1$, $u(1, \theta) = \sin(2\theta)$ using superposition.

Exercise 10.7.10: Carry out the computation we needed in the separation of variables and solve $r^2 R'' + rR' - n^2 R = 0$, for $n = 0, 1, 2, 3, \dots$

Exercise 10.7.11 (challenging): Derive the series solution to the Dirichlet problem if the region is a circle of radius ρ rather than 1. That is, solve $\Delta u = 0$, $u(\rho, \theta) = g(\theta)$.

Exercise 10.7.12 (challenging):

- a) Find the solution for $\Delta u = 0$, $u(1, \theta) = x^2y^3 + 5x^2$. Write the answer in Cartesian coordinates.
- b) Now solve $\Delta u = 0$, $u(1, \theta) = x^k y^\ell$. Write the solution in Cartesian coordinates.
- c) Suppose you have a polynomial $P(x, y) = \sum_{j=0}^m \sum_{k=0}^n c_{j,k} x^j y^k$, solve $\Delta u = 0$, $u(1, \theta) = P(x, y)$ (that is, write down the formula for the answer). Write the answer in Cartesian coordinates.

Notice the answer is again a polynomial in x and y . See also [Exercise 10.7.8](#).

Exercise 10.7.13 (challenging):* Derive the Poisson kernel solution if the region is a circle of radius ρ rather than 1. That is, solve $\Delta u = 0$, $u(\rho, \theta) = g(\theta)$.

Chapter 11

Fourier transform

11.1 Fourier Integrals

Learning Objectives

After this section, you will be able to:

- TBW

11.1.1 From Series to Integrals

We have seen how we can use Fourier series to represent and work with periodic functions. We could also use these on functions defined on finite intervals by first extending them to be periodic and then computing the sine or cosine series, depending on the problem at hand. These representations were very useful, so it would be great if there was a way to do the same thing for non-periodic functions that are defined on the entire real line.

Consider a function f that is defined on the whole real line and is not necessarily periodic. Can we do anything in terms of Fourier series? The function f isn't periodic, but we can restrict the function to an interval from $(-r, r)$, and then apply normal Fourier series representations to the restricted function. This gives that, on the interval $(-r, r)$

$$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{r}x\right) + b_n \sin\left(\frac{n\pi}{r}x\right)$$

where the coefficients are given by

$$\begin{aligned} a_0 &= \frac{1}{2r} \int_{-r}^r f(t) dt \\ a_n &= \frac{1}{r} \int_{-r}^r f(t) \cos\left(\frac{n\pi}{r}t\right) dt \\ b_n &= \frac{1}{r} \int_{-r}^r f(t) \sin\left(\frac{n\pi}{r}t\right) dt \end{aligned}$$

We can write combine the coefficients into the expression for f to get that, on this interval

$$\begin{aligned} f(x) &= \left(\frac{1}{2r} \int_{-r}^r f(t) dt \right) + \sum_{n=1}^{\infty} \left(\frac{1}{r} \int_{-r}^r f(t) \cos \left(\frac{n\pi}{r} t \right) dt \right) \cos \left(\frac{n\pi}{r} x \right) \\ &\quad + \left(\frac{1}{r} \int_{-r}^r f(t) \sin \left(\frac{n\pi}{r} t \right) dt \right) \sin \left(\frac{n\pi}{r} x \right). \end{aligned}$$

Defining $\alpha_n = \frac{n\pi}{r}$, we have that

$$\Delta\alpha = \alpha_{n+1} - \alpha_n = \frac{\pi}{r}.$$

We can then rewrite the expression for $f(x)$ as

$$\begin{aligned} f(x) &= \left(\frac{1}{2\pi} \int_{-r}^r f(t) dt \right) \Delta\alpha + \frac{1}{\pi} \sum_{n=1}^{\infty} \left(\int_{-r}^r f(t) \cos(\alpha_n t) dt \right) \cos(\alpha_n x) \Delta\alpha \\ &\quad + \left(\int_{-r}^r f(t) \sin(\alpha_n t) dt \right) \sin(\alpha_n x) \Delta\alpha. \end{aligned}$$

Now, we don't want this expression to just hold on $(-r, r)$, but on the whole real line, so we are going to take the limit of this expression as $r \rightarrow \infty$. For the integrals defining the Fourier coefficients, these become integrals from $-\infty$ to ∞ , provided the integrals converge. This expression is then a sum of terms, each multiplied by a $\Delta\alpha$. Since $\Delta\alpha = \frac{\pi}{r}$ and we are sending $r \rightarrow \infty$, this means that $\Delta\alpha \rightarrow 0$. If $\int_{-\infty}^{\infty} f(t) dt$ is finite, the first term above goes to zero, and the rest of the expression looks like

$$\frac{1}{\pi} \sum_{n=1}^{\infty} F(\alpha_n) \Delta\alpha$$

which, as $\Delta\alpha \rightarrow 0$, looks like Riemann Sum representation of an integral. Since, as $r \rightarrow \infty$, the range of α Thus, the expression here, in this limit, looks like

$$f(x) = \int_0^{\infty} \left(\int_{-\infty}^{\infty} f(t) \cos(\alpha t) dt \right) \cos(\alpha x) + \left(\int_{-\infty}^{\infty} f(t) \sin(\alpha t) dt \right) \sin(\alpha x) d\alpha.$$

This gives rise to the following definitions.

Definition 11.1.1

The *Fourier Integral* of a function f defined on $(-\infty, \infty)$ is given by

$$f(x) = \int_0^{\infty} A(\alpha) \cos(\alpha x) + B(\alpha) \sin(\alpha x) d\alpha$$

where

$$\begin{aligned} A(\alpha) &= \int_{-\infty}^{\infty} f(t) \cos(\alpha t) dt \\ B(\alpha) &= \int_{-\infty}^{\infty} f(t) \sin(\alpha t) dt. \end{aligned}$$

As with Fourier series, an important question to ask is when these integrals give rise to a proper representation of the function f . The results are a similar to, but a little more restrictive than, the requirements for Fourier series.

Theorem 11.1.1 (Convergence of Fourier Integrals)

[thm:fintConverge] Let f and f' be piecewise continuous on every finite interval and let f be absolutely integrable on $(-\infty, \infty)$, that is,

$$\int_{-\infty}^{\infty} |f(x)| dx \text{ converges.}$$

Then, the Fourier integral of f converges to $f(x)$ at every point x where f is continuous. If f is discontinuous at x , then the integral converges to the average

$$\frac{f(x+) + f(x-)}{2}$$

where

$$f(x+) = \lim_{y \rightarrow x^+} f(y) \quad f(x-) = \lim_{y \rightarrow x^-} f(y).$$

Example 11.1.1: Find the Fourier integral representation of the function

$$f(x) = \begin{cases} 1 & -1 < x < 2 \\ 0 & x < -1 \text{ or } x > 2 \end{cases}.$$

Solution: The two functions that we need to compute are

$$A(\alpha) = \int_{-\infty}^{\infty} f(t) \cos(\alpha t) dt \quad B(\alpha) = \int_{-\infty}^{\infty} f(t) \sin(\alpha t) dt.$$

Since $f(t)$ is zero outside of the interval from -1 to 2 , and 1 on that interval, both of these integral reduce to

$$A(\alpha) = \int_{-1}^2 \cos(\alpha t) dt \quad B(\alpha) = \int_{-1}^2 \sin(\alpha t) dt.$$

These can be directly computed as

$$A(\alpha) = \frac{1}{\alpha} \sin(\alpha t) \Big|_{-1}^2 \quad B(\alpha) = -\frac{1}{\alpha} \cos(\alpha t) \Big|_{-1}^2,$$

or

$$A(\alpha) = \frac{\sin(2\alpha) - \sin(-\alpha)}{\alpha} \quad B(\alpha) = \frac{\cos(-\alpha) - \cos(2\alpha)}{\alpha}.$$

Using the odd and even nature of sine and cosine, these become

$$A(\alpha) = \frac{\sin(2\alpha) + \sin(\alpha)}{\alpha} \quad B(\alpha) = \frac{\cos(\alpha) - \cos(2\alpha)}{\alpha}.$$

Then, the Fourier integral representation of f is

$$\begin{aligned} f(x) &= \frac{1}{\pi} \int_0^\infty A(\alpha) \cos(\alpha x) + B(\alpha) \sin(\alpha x) \, d\alpha \\ &= \frac{1}{\pi} \int_0^\infty \frac{\sin(2\alpha) + \sin(\alpha)}{\alpha} \cos(\alpha x) + \frac{\cos(\alpha) - \cos(2\alpha)}{\alpha} \sin(\alpha x) \, d\alpha \\ &= \frac{1}{\pi} \int_0^\infty \frac{1}{\alpha} [\sin(2\alpha) \cos(\alpha x) + \sin(\alpha) \cos(\alpha x) + \cos(\alpha) \sin(\alpha x) - \cos(2\alpha) \sin(\alpha x)] \, d\alpha. \end{aligned}$$

Then, using the identities

$$\sin(A) \cos(B) + \sin(B) \cos(A) = \sin(A + B)$$

and

$$\sin(A) \cos(B) - \sin(B) \cos(A) = \sin(A - B)$$

gives

$$f(x) = \frac{1}{\pi} \int_0^\infty \frac{1}{\alpha} [\sin(\alpha(2-x)) + \sin(\alpha(1+x))] \, d\alpha.$$

Finally, the identity

$$\sin(A) + \sin(B) = 2 \sin \frac{A+B}{2} \cos \frac{A-B}{2}$$

gives the final representation as

$$f(x) = \frac{2}{\pi} \int_0^\infty \frac{1}{\alpha} \left[\sin \frac{3\alpha}{2} \cos \frac{\alpha(1-2x)}{2} \right] \, d\alpha.$$

—

We can also use these representations to compute integrals. For example Theorem ?? says that if we plug in $x = \frac{1}{2}$, the integral will converge to $f(\frac{1}{2})$, which we know equals 1. This gives that

$$1 = \frac{2}{\pi} \int_0^\infty \frac{1}{\alpha} \left[\sin \frac{3\alpha}{2} \cos \frac{\alpha(0)}{2} \right] \, d\alpha,$$

or

$$\frac{\pi}{2} = \int_0^\infty \frac{\sin(\frac{3}{2}\alpha)}{\alpha} \, d\alpha.$$

Exercise 11.1.1: Find the Fourier integral representation of the function

$$f(x) = \begin{cases} 1 & 0 < x < 2 \\ 0 & x < 0 \text{ or } x > 2 \end{cases}$$

and use this representation to find the value of $\int_0^\infty \frac{\sin(\alpha)}{\alpha}$.

11.1.2 Cosine and Sine Integrals

Just like cosine and sine series for representing even and odd periodic functions, we can also talk about cosine and sine integrals for representing even and odd functions on the entire real line. As before, these are the cosine and sine components of the fourier integrals. If $f(x)$ is an even function, then the product $f(x) \cos(\alpha x)$ is an even function and $f(x) \sin(\alpha x)$ is an odd function. In this case, we know that

$$B(\alpha) = \int_{-\infty}^{\infty} f(t) \sin(\alpha t) dt = 0$$

and that

$$A(\alpha) = \int_{-\infty}^{\infty} f(t) \cos(\alpha t) dt = 2 \int_0^{\infty} f(t) \cos(\alpha t) dt.$$

Theorem 11.1.2 (Cosine and Sine Integrals)

The Fourier integral of an even function f on $(-\infty, \infty)$ is the *cosine integral*

$$f(x) = \frac{1}{\pi} \int_0^{\infty} A(\alpha) \cos(\alpha x) d\alpha$$

where

$$A(\alpha) = 2 \int_0^{\infty} f(t) \cos(\alpha t) dt.$$

The Fourier integral of an odd function f on $(-\infty, \infty)$ is the *sine integral*

$$f(x) = \frac{1}{\pi} \int_0^{\infty} B(\alpha) \sin(\alpha x) d\alpha$$

where

$$B(\alpha) = 2 \int_0^{\infty} f(t) \sin(\alpha t) dt.$$

Example 11.1.2: Compute the Fourier integral representation of the function

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ -1 & -1 < x < 0 \\ 0 & x = 0, |x| > 1 \end{cases}.$$

Solution: This is an odd function, so we know that the $A(\alpha)$ function will be zero and we can compute $B(\alpha)$ by

$$B(\alpha) = 2 \int_0^{\infty} f(t) \sin(\alpha t) dt = 2 \int_0^1 \sin(\alpha t) dt = -\frac{2}{\alpha} (\cos(\alpha) - 1).$$

Thus, the Fourier integral representation for this function f is

$$f(x) = \frac{1}{\pi} \int_0^{\infty} \frac{2}{\alpha} (1 - \cos(\alpha)) \sin(\alpha x) d\alpha,$$

or

$$f(x) = \frac{2}{\pi} \int_0^\infty \frac{(1 - \cos(\alpha)) \sin(\alpha x)}{\alpha} d\alpha.$$

□

Just like for Fourier series, we can use sine and cosine Fourier integrals to represent the odd and even extensions of functions that are defined on $(0, \infty)$. If f is defined on $(0, \infty)$, then with

$$A(\alpha) = 2 \int_0^\infty f(t) \cos(\alpha t) dt \quad B(\alpha) = 2 \int_0^\infty f(t) \sin(\alpha t) dt,$$

the function

$$F(x) = \frac{1}{\pi} \int_0^\infty A(\alpha) \cos(\alpha x) d\alpha$$

is the Fourier integral representation of the even extension of f on $(-\infty, \infty)$ and

$$G(x) = \frac{1}{\pi} \int_0^\infty B(\alpha) \sin(\alpha x) d\alpha$$

is the Fourier integral representation of the odd extension of f on $(-\infty, \infty)$.

Exercise 11.1.2: Show that the representation of e^{-x} on $(0, \infty)$ as a cosine integral is

$$f(x) = \frac{2}{\pi} \int_0^\infty \frac{\cos(\alpha x)}{1 + \alpha^2} d\alpha$$

and the representation as a sine integral is

$$f(x) = \frac{2}{\pi} \int_0^\infty \frac{\alpha \sin(\alpha x)}{1 + \alpha^2} d\alpha.$$

11.1.3 Complex Form

Just like Fourier series had a complex form, Fourier Integrals can also be written in a complex form. We want to use Euler's formula

$$e^{ix} = \cos(x) + i \sin(x)$$

to rearrange and simplify our expression. First we write f using its Fourier integral representation and combine terms

$$\begin{aligned} f(x) &= \frac{1}{\pi} \int_0^\infty \left(\int_{-\infty}^\infty f(t) \cos(\alpha t) dt \right) \cos(\alpha x) + \left(\int_{-\infty}^\infty f(t) \sin(\alpha t) dt \right) \sin(\alpha x) d\alpha \\ &= \frac{1}{\pi} \int_0^\infty \int_{-\infty}^\infty f(t) [\cos(\alpha t) \cos(\alpha x) + \sin(\alpha t) \sin(\alpha x)] dt d\alpha \\ &= \frac{1}{\pi} \int_0^\infty \int_{-\infty}^\infty f(t) \cos(\alpha(t - x)) dt d\alpha \end{aligned}$$

Now, because $f(t) \cos(\alpha(t - x))$ is an even function of α (since cosine is an even function), we can extend the interval from $-\infty$ to ∞ on the outside integral and divide the expression by 2, to give

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(t) \cos(\alpha(t - x)) dt d\alpha.$$

Next, since $f(t) \sin(\alpha(t - x))$ is an odd function of α , we know that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(t) \sin(\alpha(t - x)) dt d\alpha = 0$$

so that we can add i times this term to our previous expression for f to get that

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(t) [\cos(\alpha(t - x)) + i \sin(\alpha(t - x))] dt d\alpha.$$

Then, Euler's formula gives that

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(t) e^{i\alpha(t-x)} dt d\alpha = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(t) e^{i\alpha(t)} dt \right) e^{-i\alpha x} d\alpha.$$

By rearranging this expression, we get the complex form of the Fourier integral.

Theorem 11.1.3 (Complex Fourier Integrals)

The complex form of the Fourier integral representation of a function f is given by

$$f(x) = \int_{-\infty}^{\infty} C(\alpha) e^{-i\alpha x} d\alpha$$

where

$$C(\alpha) = \int_{-\infty}^{\infty} f(t) e^{i\alpha t} dt.$$

11.1.4 Use of Computers

For all of these integral representations, numerical and graphical representations can be used to approximate them. For example, for a function f , the cosine integral of f is given by

$$A(\alpha) = \int_0^{\infty} f(t) \cos(\alpha t) dt$$

and we can represent the even extension of f on $(0, \infty)$ by

$$F(x) = \frac{2}{\pi} \int_0^{\infty} A(\alpha) \cos(\alpha x) d\alpha.$$

If we can work out the function $A(\alpha)$ analytically, then we can use the function

$$F_r(x) = \frac{2}{\pi} \int_0^r A(\alpha) \cos(\alpha x) d\alpha$$

as an approximation to the function $f(x)$, similar to how we can use partial sums of a Fourier series to approximation a function. For example, for the function $f(x) = e^{-2x}$ on $(0, \infty)$, we can compute that the cosine integral is

$$A(\alpha) = \frac{2}{4 + \alpha^2}$$

and the sine integral is

$$B(\alpha) = \frac{\alpha}{4 + \alpha^2}.$$

Therefore, we can approximate the even extension of f by

$$F_r(x) = \frac{2}{\pi} \int_0^r \frac{2 \cos(\alpha x)}{4 + \alpha^2} d\alpha$$

and the odd extension of f by

$$G_r(x) = \frac{2}{\pi} \int_0^r \frac{\alpha \sin(\alpha x)}{4 + \alpha^2} d\alpha$$

for large enough r , since

$$\lim_{r \rightarrow \infty} F_r(x) = f(x) \quad \lim_{r \rightarrow \infty} G_r(x) = f(x)$$

on $(0, \infty)$ with the functions F_r and G_r being even and odd respectively. [Figure 11.1](#) on the next page through [Figure 11.4](#) on the facing page illustrate the even and odd extensions of the function $f(x) = e^{-2x}$ with the approximations F_r and G_r with $r = 5$ and $r = 30$, generated using Matlab.

11.1.5 Exercises

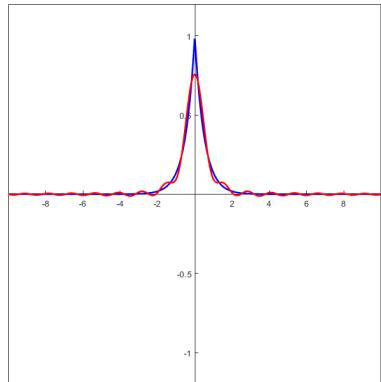


Figure 11.1: Even extension of e^{-2x} with $r = 5$ approximation.

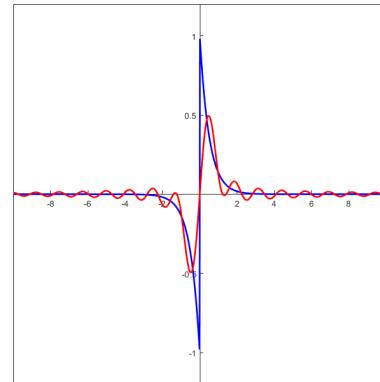


Figure 11.2: Odd extension of e^{-2x} with $r = 5$ approximation.

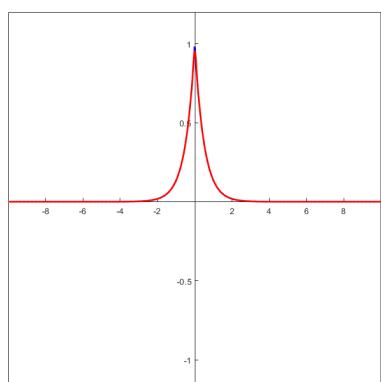


Figure 11.3: Even extension of e^{-2x} with $r = 30$ approximation.

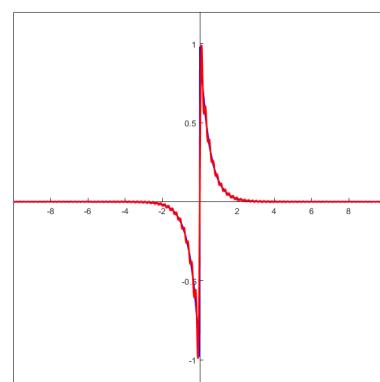


Figure 11.4: Odd extension of e^{-2x} with $r = 30$ approximation.

11.2 Fourier Transform

Learning Objectives

After this section, you will be able to:

- TBW

11.2.1 Integral Transform Pairs

A lot of the work that we did in the last section with Fourier integrals looks a lot like Laplace transforms from [chapter 6](#). Using an integral, we transform a function of one variable (x in this case, t in Laplace transforms) to another variable (α here, s for Laplace transforms) and can use it to represent the function. The main difference here is that all of the integrals in [§ 11.1](#) have an easy way to go back to the original function (using another integral), while to “go back” with Laplace transforms required writing expressions in a specific form and using a table.

The Laplace transform (and the Fourier transform that will be defined shortly) is an example of a more general class of operations called **integral transforms**, and for all of these transforms, both the forward and inverse transforms can be written as an integral. However, we never used the integral inverse transform for the Laplace transform because it was not simple to compute. For the Laplace transform, we have the forward transform

$$\mathcal{L}\{f(t)\} = \int_0^\infty e^{-st} f(t) dt = F(s)$$

and the inverse transform is defined by

$$\mathcal{L}^{-1}\{F(s)\} = \frac{1}{2\pi i} \int_{\gamma-i\infty}^{\gamma+i\infty} e^{st} F(s) ds = f(t).$$

This last integral looks strange; this is what is called a **contour integral**. It requires complex numbers to compute, and is well beyond the scope of this book and course. For this reason, we have always used a table for computing the inverse Laplace transform of a function.

These transforms always come in **transform pairs**. If $f(x)$ is transformed into $F(\alpha)$ by an integral transform

$$F(\alpha) = \int_a^b f(x) K(\alpha, x) dx,$$

then the inverse transform can also be represented by an integral transform

$$f(x) = \int_c^b F(\alpha) H(\alpha, x) d\alpha.$$

The functions K and H are called the **kernel** of these transforms. For instance, in the case of the Laplace transform, we have

$$K(s, t) = e^{-st} \quad H(s, t) = \frac{e^{st}}{2\pi i}.$$

11.2.2 Fourier Transform

The integrals that we worked with in the previous section give rise to three different integral transforms, all referred to as Fourier transform pairs.

Definition 11.2.1

The **Fourier Transform** of a function f defined on $(-\infty, \infty)$ is

$$\mathcal{F}\{f(x)\} = \int_{-\infty}^{\infty} f(x)e^{i\alpha x} dx = F(\alpha)$$

and the corresponding inverse transform is define by

$$\mathcal{F}^{-1}\{F(\alpha)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\alpha)e^{-i\alpha x} d\alpha = f(x).$$

The **Fourier Cosine Transform** of a function f defined on $(0, \infty)$ is

$$\mathcal{F}_c\{f(x)\} = \int_0^{\infty} f(x) \cos(\alpha x) dx = F(\alpha)$$

and the corresponding inverse transform is define by

$$\mathcal{F}_c^{-1}\{F(\alpha)\} = \frac{2}{\pi} \int_0^{\infty} F(\alpha) \cos(\alpha x) d\alpha = f(x).$$

The **Fourier Sine Transform** of a function f defined on $(0, \infty)$ is

$$\mathcal{F}_s\{f(x)\} = \int_0^{\infty} f(x) \sin(\alpha x) dx = F(\alpha)$$

and the corresponding inverse transform is define by

$$\mathcal{F}_s^{-1}\{F(\alpha)\} = \frac{2}{\pi} \int_0^{\infty} F(\alpha) \sin(\alpha x) d\alpha = f(x).$$

Properties of the Transform

Like we saw for Fourier integrals in the previous section, the conditions for existence of Fourier transforms is more restrictive than for Laplace transforms. For example, with the constant function 1, $\mathcal{F}\{1\}$, $\mathcal{F}_c\{1\}$, and $\mathcal{F}_s\{1\}$ all do not exist. A condition that is sufficient for existence of these transforms is that f be absolutely integrable (that is $\int_{-\infty}^{\infty} |f(x)| dx$ exists), and f and f' are piecewise continuous on every finite interval.

Next, we want to see how these transforms behave under a variety of situations. Firstly, these transforms, like all other integral transforms, are linear. This means that if the transforms of f and g exist, then for any two constants c_1 and c_2 ,

$$\mathcal{F}\{c_1 f(x) + c_2 g(x)\} = c_1 \mathcal{F}\{f(x)\} + c_2 \mathcal{F}\{g(x)\}.$$

This works for any of the transforms given above.

The other main consideration is how these transforms behave with respect to differentiation. The main use of the Laplace transform was that it could change differentiation into multiplication, so we could use algebraic manipulation to solve differential equations. The Fourier transform works in a very similar way.

Assume that f is continuous, absolutely integrable on $(-\infty, \infty)$, and f' is piecewise continuous on every finite interval. The conditions above tell us that we must have $f(x) \rightarrow 0$ as $x \rightarrow \pm\infty$. With this, integration by parts tells us that

$$\begin{aligned}\mathcal{F}\{f'(x)\} &= \int_{-\infty}^{\infty} f'(x)e^{i\alpha x} dx \\ &= f(x)e^{i\alpha x} \Big|_{-\infty}^{\infty} -i\alpha \int_{-\infty}^{\infty} f(x)e^{i\alpha x} dx \\ &= -i\alpha \int_{-\infty}^{\infty} f(x)e^{i\alpha x} dx.\end{aligned}$$

Therefore, we see that

$$\mathcal{F}\{f'(x)\} = (-i\alpha)F(\alpha)$$

where $F(\alpha)$ is the Fourier transform of the function $f(x)$.

By the same process, if we also assume that f' is continuous, f'' piecewise continuous on every finite interval, and $f'(x) \rightarrow 0$ as $x \rightarrow \pm\infty$, then we have that

$$\mathcal{F}\{f''(x)\} = (-i\alpha)^2 \mathcal{F}\{f(x)\} = -\alpha^2 F(\alpha).$$

Repeating this under similar assumptions gives that

$$\mathcal{F}\{f^{(n)}(x)\} = (-i\alpha)^n \mathcal{F}\{f(x)\} = (-i\alpha)^n F(\alpha)$$

for any number of derivatives n .

However, the sine and cosine transforms are suitable for representing the first (or any odd) derivative of a function. The main issue is that the integration by parts differentiates the other part of the expression, which swaps sine and cosine, but preserves exponential functions. If the same process is carried out (under the same assumptions) for the sine and cosine transforms, the result is that

$$\mathcal{F}_c\{f'(x)\} = \alpha \mathcal{F}_s\{f(x)\} - f(0) \quad \mathcal{F}_s\{f'(x)\} = -\alpha \mathcal{F}_c\{f(x)\}.$$

Since the same transform does not appear on both sides of the expression, we are not able to use these to convert differentiation into algebra in order to solve the equation.

11.2.3 Fourier Sine and Cosine Transforms

So, what can we use the sine and cosine transforms for? It turns out they are really useful for situations where we have an even number of derivatives. Suppose that both f and f' are

continuous, absolutely integrable on the interval $[0, \infty)$, and $f \rightarrow 0, f' \rightarrow 0$ as $x \rightarrow \infty$. Then the Fourier cosine transform of f'' is, using integration by parts twice,

$$\begin{aligned}\mathcal{F}_c\{f''(x)\} &= \int_0^\infty f''(x) \cos(\alpha x) dx \\ &= f'(x) \cos(\alpha x) \Big|_0^\infty + \alpha \int_0^\infty \sin(\alpha x) dx \\ &= -f'(0) + \alpha \left[f(x) \sin(\alpha x) \Big|_0^\infty - \alpha \int_0^\infty f(x) \cos(\alpha x) dx \right] \\ &= -f'(0) - \alpha^2 \mathcal{F}_c\{f(x)\}.\end{aligned}$$

Thus, we have that

$$\mathcal{F}_c\{f''(x)\} = -\alpha^2 F(\alpha) - f'(0)$$

where $F(\alpha) = \mathcal{F}_c\{f(x)\}$.

The same process can be used for the Fourier sine transform, giving that

$$\mathcal{F}_s\{f''(x)\} = -\alpha^2 F(\alpha) + \alpha f(0)$$

where $F(\alpha) = \mathcal{F}_s\{f(x)\}$.

So, both of these work in similar ways, differing in whether the expression involves $f(0)$ or $f'(0)$.

11.2.4 Applications

As with Laplace transforms, we want to use Fourier transforms to solve differential equations problems, and in this case, we want to do use this for boundary value problems. How do we choose which of the three options to use? The domain of the variable that we want to transform in is the first big indicator. If the domain is $(-\infty, \infty)$, then the complex exponential Fourier transform is the way to go. If the domain is $[0, \infty)$, then either the cosine or sine transform will work better, as these are defined on these domains. In choosing between the two of them, we should think about whether having $f(0)$ or $f'(0)$ in the expression will be more convenient. The following examples will illustrate these ideas.

Example 11.2.1: Solve the heat equation $k \frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}$ on $-\infty < x < \infty, t > 0$ with the initial condition

$$u(x, 0) = f(x) = \begin{cases} 10 & |x| < r \\ 0 & x > r \end{cases}$$

for some number r .

Solution: Since the x domain is $(-\infty, \infty)$, we should use the normal complex Fourier transform in the x variable and see how this changes the equation. Thus, we define

$$U(\alpha, t) = \mathcal{F}\{u(x, t)\} = \int_{-\infty}^\infty u(x, t) e^{i\alpha x} dx$$

and apply the Fourier transform to the differential equation and initial condition. When we do this

$$k \frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}$$

becomes

$$k(-i\alpha)^2 U(\alpha, t) = \frac{\partial U}{\partial t}$$

and the initial condition becomes

$$U(\alpha, 0) = \mathcal{F}\{u(x, 0)\} = \int_{-\infty}^{\infty} f(x) e^{i\alpha x} dx$$

which we can evaluate to be

$$\int_{-r}^r 10e^{i\alpha x} dx = \frac{10}{i\alpha} e^{i\alpha x} \Big|_{-r}^r = \frac{20}{\alpha} \frac{e^{i\alpha r} - e^{-i\alpha r}}{2i} = \frac{20}{\alpha} \sin(\alpha r).$$

The last simplification here comes from Euler's formula, since

$$e^{i\alpha r} = \cos(\alpha r) + i \sin(\alpha r) \quad e^{-i\alpha r} = \cos(\alpha r) - i \sin(\alpha r).$$

This means we are trying to solve

$$\frac{\partial U}{\partial t} = -k\alpha^2 U(\alpha, t) \quad U(\alpha, 0) = \frac{20}{\alpha} \sin(\alpha r).$$

For each α , this is a normal ordinary differential equation in t . The general solution to this differential equation is

$$U(\alpha, t) = C(\alpha) e^{-k\alpha^2 t}$$

and based on the value at $U(\alpha, 0)$, the solution we want is

$$U(\alpha, t) = \frac{20}{\alpha} \sin(\alpha r) e^{-k\alpha^2 t}.$$

To get the solution u , we then want to take the inverse Fourier transform, which gives

$$u(x, t) = \mathcal{F}^{-1}\{U(\alpha, t)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{20}{\alpha} \sin(\alpha r) e^{-k\alpha^2 t} e^{-i\alpha x} d\alpha.$$

We can simplify this a bit further, but we won't be able to evaluate the integral. Using the fact that $e^{-i\alpha x} = \cos(\alpha x) - i \sin(\alpha x)$ and that

$$\frac{20}{\alpha} \sin(\alpha r) e^{-k\alpha^2 t} \sin(\alpha x)$$

is an odd function, so the integral will be zero. Thus, we get that

$$u(x, t) = \frac{10}{\pi} \int_{-\infty}^{\infty} \frac{\sin(\alpha r) \cos(\alpha x)}{\alpha} d\alpha.$$

□

Example 11.2.2: Solve Laplace's equation on a semi-infinite plate,

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= 0 \quad 0 < x < \pi, \quad y > 0 \\ u(0, y) &= 0, \quad u(\pi, y) = e^{-2y}, \quad y > 0 \\ u(x, 0) &= 0, \quad 0 < x < \pi. \end{aligned}$$

Solution: In this example, the x domain is finite, $(0, \pi)$, but the y domain is semi-infinite, since it is $(0, \infty)$. This means that we'll need to use either the Fourier cosine or Fourier sine transform in y in order to solve this problem. Let $U(x, \alpha)$ be the transformed function, and let's see which transform is more helpful here. The important term here is the $\frac{\partial^2 u}{\partial y^2}$ term. If we use the sine transform, we get that

$$\mathcal{F}_s\left\{\frac{\partial^2 u}{\partial y^2}\right\} = -\alpha^2 U(x, \alpha) + \alpha u(x, 0)$$

and for the cosine transform, we get that

$$\mathcal{F}_c\left\{\frac{\partial^2 u}{\partial y^2}\right\} = -\alpha^2 U(x, \alpha) + \frac{\partial u}{\partial y}(x, 0),$$

since the f and f' terms in the previous definitions become values of u and $\frac{\partial u}{\partial y}$ at $(x, 0)$. Since the problem tells us that $u(x, 0) = 0$, it is much more convenient to use the sine transform, because that term appears in the computation, and will go to zero. Therefore, using the sine transform, we get that

$$\frac{\partial^2 U}{\partial x^2} + -\alpha^2 U(x, \alpha) + 0 = 0$$

with boundary conditions

$$U(0, \alpha) = \mathcal{F}_s 0 = 0$$

and

$$U(\pi, \alpha) = \mathcal{F}_s\{e^{-2y}\} = \int_0^\infty e^{-2y} \sin(\alpha x) dx,$$

which, after two integrations by parts, evaluates to

$$U(\pi, \alpha) = \frac{\alpha}{4 + \alpha^2}.$$

Thus, for each α , we have the second order differential equation

$$\frac{\partial^2 U}{\partial x^2} - \alpha^2 U(x, \alpha) = 0 \quad U(0, \alpha) = 0, \quad U(\pi, \alpha) = \frac{\alpha}{4 + \alpha^2}.$$

The general solution to the differential equation is

$$U(x, \alpha) = C_1 e^{\alpha x} + C_2 e^{-\alpha x}$$

or, in a more convenient form for these boundary conditions,

$$U(x, \alpha) = C_1 \cosh(\alpha x) + C_2 \sinh(\alpha x).$$

The fact that $U(0, \alpha) = 0$ gives that $C_1 = 0$, and the second condition tells us that

$$C_2 = \frac{\alpha}{(4 + \alpha^2) \sinh(\alpha\pi)}.$$

Thus, the particular solution that we want is

$$U(x, \alpha) = \frac{\alpha \sinh(\alpha x)}{(4 + \alpha^2) \sinh(\alpha\pi)}.$$

By taking the inverse Fourier sine transform, we get that

$$u(x, y) = \frac{2}{\pi} \int_0^\infty \frac{\alpha \sinh(\alpha x)}{(4 + \alpha^2) \sinh(\alpha\pi)} \sin(\alpha x) d\alpha$$

□

Exercise 11.2.1: Repeat the last example with the final boundary condition changed to $\frac{\partial u}{\partial y}(x, 0) = 0$ on $0 < x < \pi$. You will want to use the cosine transform for this problem.

11.2.5 Exercises

Appendix A

Introduction to MATLAB

This document is meant to provide a review of some of the main skills and techniques in MATLAB that are necessary to complete the various MATLAB assignments throughout the course. In addition, these skills will be useful when attempting to use MATLAB, both for illustrating problems in differential equations and for solving other types of problems that can be analyzed using this software.

A.1 The MATLAB Interface

There are many components to the MATLAB interface, and the way that the window is organized can be fully customized. There are four main components of this interface.

1. Current Folder window. This shows the current folder in which MATLAB is running. This determines what files that MATLAB currently has access to and what functions and methods can be called.
2. Editor window. This is the main code-editing window, where script files can be written, edited, saved, and run.
3. Command window. This is where individual lines of code can be entered to see how they work.
4. Workspace window. This shows a list of all variables that currently exist, as well as their values or sizes.

All four of these components are very useful in organizing thoughts and programming practices while using MATLAB. Both the Default layout and Two-Column layout (as of MATLAB R2019b) contain all four of these windows in different locations. Either of these will work for programming in MATLAB, as well as any modifications of them. The current format can be saved using Layout - Save Layout if needed.

A.1.1 File Structure

The main type of file used in MATLAB is the Script file. These are saved as ‘*.m’ files and can represent both stand-alone executable files and functions that can be called from

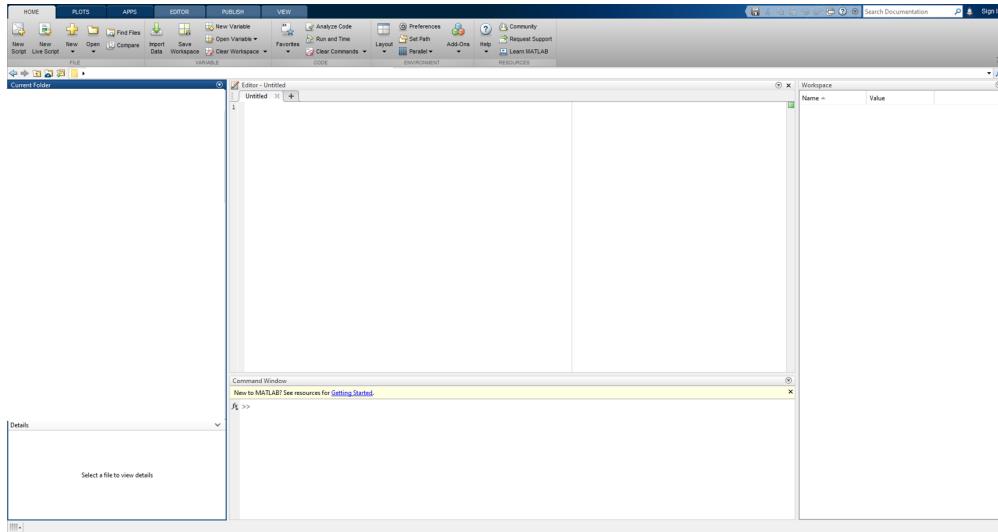


Figure A.1: The default layout provided by MATLAB.

other scripts. For running simple, one-line expressions or debugging code, the Command Window and the command line prompt can be useful. However, for anything more involved and complicated than that, the script editor should be used instead.

In writing a script file or using the Command window, the Current Folder window shows all of the files in the current directory. These are all of the files that MATLAB has access to while running a MATLAB file that it saved in that folder. This means that if a script wants to call a method, it either needs to be a built-in method or a function file that is contained within the same script file or the Current Folder. For more information about writing functions, see Section A.4.

To use script files, multiple lines of code can be entered in a row, and MATLAB will execute them in sequence when the “Run” button is clicked. This button is in the “Editor” tab at the top of the screen.

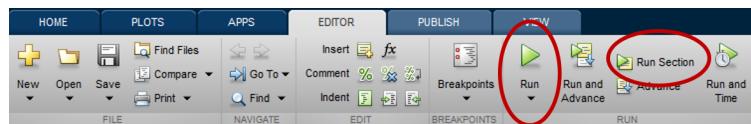


Figure A.2: Location of the Run buttons on the MATLAB interface.

MATLAB Live Scripts can also be used to do very similar things, with some additional benefits. These allow the MATLAB code to be viewed side-by-side with the output, as well as an easy export to PDF functionality. These are saved as ‘*.mlx’ files. These work the same way as scripts in terms of how code is written, and allow the user to mix between text (which can be resized and formatted) and code. For more information on Live Scripts, see the website https://www.mathworks.com/help/matlab/matlab_prog/what-is-a-live-script-or-function.html.

Live Scripts also have the ability to put section breaks between different pieces of code and then run individual sections using the “Run Section” button at the top of the editor.

With Live Scripts, it is necessary to run the entire code (by clicking the run button) before exporting as a PDF in order to get the correct images and outputs in the final PDF. To export, go to Save at the top of the screen, click the down arrow under it, and select “Export to PDF” after running the code to regenerate all of the images.

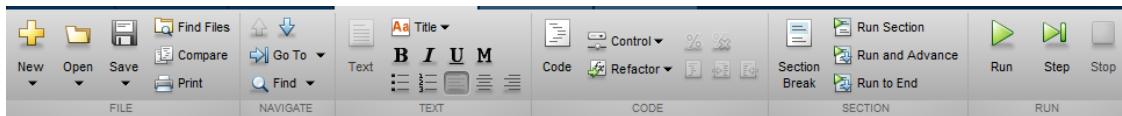


Figure A.3: Header Bar for the MATLAB Live Script Interface.

A.2 Computation in MATLAB

MATLAB can do many of the simple computational operations that would be expected from a calculator. It is easiest to see these operations by using the Command Window, but they can also be implemented in scripts if desired. Addition and subtraction work in standard ways. In the command line, typing

```
2 + 3
```

and pressing ENTER will give an output of

```
ans =
5
```

showing the answer of this computation. For any computation or line of code, putting a semi-colon (;) at the end will suppress the output, in that typing

```
2 + 3;
```

will not show any output. However, MATLAB did do the computation, which can be shown by storing this output in a variable and doing something with it later.

Multiplication and division, and by extension powers, can work differently in MATLAB. As MATLAB is built around using matrices for calculations and is optimized for this approach, the program interprets all multiplication, division, and exponentiation in terms of matrices as a default. Both components of the multiplication are simple scalars (numbers), then this is fine. The “*” symbol works for multiplication in this context:

```
>> 4*6
ans =
24
```

as well as using ‘/’ for division and ‘^’ for exponentiation. Issues may arise when the code wants to compute products or powers of multiple values at the same time. Many MATLAB

built-in functions will automatically combine multiple of the same type of calculation into a ‘vectorized’ calculation, where if the code wanted to compute the sum of two numbers a bunch of times, it would put all of these numbers into arrays and then add the two vectors together. This completes the task of adding all of the different pairs of numbers together, but saves time by not doing them all individually. This works great for addition and subtraction, because addition and subtraction of arrays or matrices is done element-wise, which is the exact operation we wanted to compute in the first place.

However, multiplication is different. Matrix multiplication is a different operation that, in particular, is not element-wise multiplication. Beyond that, even if two matrices are the same size, it is possible that their product, in the normal matrix sense, is not defined. In MATLAB, the product

```
[1 2 3] * [4 3 2];
```

will return an error because the matrices are not the correct size. From a human point of view, the output desired from this code was likely [4 6 6], the product of each term individually. To obtain this in MATLAB, we need the elementwise operations ‘.*’, ‘./’ and ‘.^’ for multiplication, division, and exponentiation, respectively. Thus, the following computations can be made in MATLAB

```
>> [1 2 3] .* [4 3 2]
ans =
    [4 6 6]
>> [1 4 6].^2
ans =
    [1 16 36]
>> [5 4 2] ./ [10 2 6]
ans =
    [0.5 2 0.3333]
```

There are many built-in functions in MATLAB that can help with computation and algebra.

- `sqrt(x)` will compute the square root of a number x .
- `exp(x)` will compute e^x for e the base of the natural logarithm, and x any number. Note that MATLAB does not know the definition of e built-in, so it will either need to be defined (using `exp(1)`) or just use `exp()` whenever it is needed.
- `abs(x)` computes the absolute value of a number x .
- `log(x)` computes the natural logarithm of a number x . The functions `log2` and `log10` compute the log base 2 and log base 10 respectively.
- Trigonometric functions can also be computed with `sin(x)`, `cos(x)`, and `tan(x)`.

A.3 Variables and Arrays

As with other programming languages, MATLAB utilizes variables to store information and use it later. The name of variables in MATLAB must start with a letter, but the rest of the name can consist of letters, digits, or underscores. Variables should be named suggestively corresponding to what this information is or the way it will be used. Variables do not need to be created in advance, they are created when something is stored in the variable by putting the name on the left side of an equals sign, with the computation that gives rise to that variable on the right. Even though the output is suppressed, the line

```
val = 2+3;
```

will store the value 5 in the variable `val`, where it can be used later. For example,

```
>> val * 4
ans =
20

>> val^2 + 2
ans =
27
```

However, trying to use a variable name without defining it first will cause MATLAB to give an error:

```
>> r
Undefined function or variable 'r'.
```

As variables do not need to be created or instantiated before they are used, any variable can store any type of information. Two of the most common ones are numbers (double precision) or strings.

```
numVar = sqrt(15);
strVar = ``Hello World!'';
```

Strings can be stored using either single or double quotes. Strings also have a lot of useful operations that can be used to make some MATLAB programs run more simply, but they are beyond the scope of this introduction. For information about what can be done with strings, see the MATLAB documentation <https://www.mathworks.com/help/matlab/ref/string.html>.

Another common variable data type that MATLAB is very comfortable with is arrays. As described previously, MATLAB defaults to matrices when considering multiplication and exponentiation operations. Arrays can be created using square brackets, with either spaces or commas between the entries.

```
A = [2,4,6];
B = [1 3 5];
```

These create horizontal arrays. Vertical arrays can also be created using semi-colons between each entry, and these can be combined with horizontal arrays to create a matrix, or rectangular array of values.

```
C = [5;7;8];
M = [1,2,3;5,6,7];
```

In these examples, A and B will be row arrays (or row vectors) with 3 elements, C will be a column vector with 3 elements, and M will be a matrix with two rows and three columns. For most situations that don't involve matrices, row and column vectors will work equivalently, so either one can be used. Once matrices are involved, it matters which one is chosen, because MATLAB will multiply matrices and vectors in the same way that would be carried out mathematically, which means the dimensions need to match.

To access elements of a matrix, parentheses are used. Unlike other programming languages, MATLAB starts indexing elements at 1, not zero. That is, with the above variables $C(2) = 7$, since 7 is the second element of the array C . In terms of accessing elements of matrices, the first index is the row and the second is the column.

```
>> M = [1,2,3;5,6,7];
>> M(1,1)
ans =
    1

>> M(1,3)
ans =
    3

>> M(2,1)
ans =
    5
```

The matrix (and vectors) do have limits on how big they are, and attempting to access an element outside of that range will cause MATLAB to give an error.

```
>> M(3,1)
Index in position 1 exceeds array bounds (must not exceed 2).
```

Among many other possible variables, another type that can be stored is a handle to a function. How to use functions will be described in Section A.4. The fact that all of these different data types can be stored in variables, with no real indication as to which type a given variable is, means it is critical to name variables carefully with what they correspond to.

A.4 Functions and Anonymous Functions

A key component to programming in MATLAB is the idea of functions. These are programming objects that will accept a number of inputs (called *arguments*) and perform a given set of operations on those arguments, returning some set of outputs back to the main program. These are mainly used to group code together that has a given purpose and can be called to carry out that purpose on a variety of outputs. An example of a built-in function like this is `sum(V)`. This function takes in a linear array and will return the number that is the sum of all of the elements in the array (if the array is multi-dimensional, it will only sum along one dimension). This is a piece of code that could be written fairly easily; it would just involve taking the array, looping through it and adding up the value at each index. However, putting it into a function allows it to be called more simply in one line, allowing the main script to focus on the task at hand.

There are two main ways that functions can be written in MATLAB. Functions can either be written at the bottom of the MATLAB script where they will be used or they can be written in their own separate script file. If written in a separate file, there can only be one function in each file, and the name of the file (once saved) must match the name given to the function. To write a function, the reserved word ‘function’ is used:

```
function [a,b] = testFunction(x, y, z)
    % Code here
end
```

Note: If this is done in a script by itself, the function line must be the first line of the code. There can be no code or comments above this line.

In this case, the function takes in three inputs and returns two outputs. When writing the code inside the function, the three inputs will be called `x`, `y`, and `z`, and in order to tell the program what to send back to wherever this function was called, those outputs should be stored in variables `a` and `b`. For example, a function that takes in three numbers and returns their sum in the first output and the product in the second would look like

```
function [a,b] = testFunction(x, y, z)
    a = x+y+z;
    b = x*y*z;
end
```

and that would work just fine. However, if any other MATLAB methods were going to use this function, there is a chance they would try to pass in array inputs. If so, then there would be an error in computing `b`, because those products would not be defined. The easiest way to fix this would be to use element-wise products, giving a function that looks like

```
function [a,b] = testFunction(x, y, z)
    a = x+y+z;
    b = x.*y.*z;
end
```

These functions can be as complicated as necessary, including graphs, loops, calls to other functions, and many different components. However, if the function needed is a simple mathematical function, then this can be written in a shorter way with anonymous functions. For example, if the function $f(x,y) = x^2 + 4xy + y^2$ needed to be coded, it could be written as

```
f = @(x,y) x.^2 + 4.*x.*y + y.^2;
```

and this will now make `f` a handle to the function that does exactly what is desired. If a later line of code is

```
>> f(2,1)
ans =
    13
```

the function value will be computed at the desired point. Notice the use of element-wise operations again in this function definition to ensure that it will also work on array inputs. This works for these simple kinds of functions, and can be easier than adding an entire new function to the script file.

Overall, the following two function definitions are *almost* equivalent.

```
fShort = @(x,y) x.^2 + y.^2;
```

```
function z = fLong(x,y)
z = x.^2 + y.^2;
end
```

The only difference arises when trying to use these functions in built-in or written methods that require a handle to a function. The '@' symbol at the beginning of the anonymous function indicates that the thing being defined (`fShort`) is a handle to a function that takes two inputs and computes an output from it. On the other hand, the definition of `fLong` is a function that does this, and is not a handle to that function. To fix this, an '@' symbol needs to be put in-front of `fLong` before using it in one of these methods. As an example `ode45` is a method that numerically computes the solution to a differential equation, and it requires a function handle in the first argument. So, the code

```
ode45(fShort, [0, 3], 1)
```

runs fine. However,

```
ode45(fLong, [0, 3], 1)
```

throws an error about there being not enough inputs for `fLong`. This is because whenever MATLAB sees `fLong`, it is expecting to see two inputs next to it. This is not the case for `fShort` because of the way it was defined. To remedy this, the code needs to be written

```
ode45(@fLong, [0, 3], 1)
```

and then it will execute the same as the first line.

With any of these functions, it is possible to restrict variables and get new functions. This can be fairly easily done with the same setup as for anonymous functions. The line of code

```
fNew = @(y) fShort(1,y)
```

will create a new handle for a function of one variable that is `fShort` when the x value is fixed to be 1. The exact same code will work for `fLong` as you are giving it two inputs.

A.5 Loops and Branching Statements

The code written in a MATLAB script will always proceed in order from one line to the next unless there is some alteration to the flow using loops or branching (if) statements.

A.5.1 For Loops

For loops are a form of iterative programming, where MATLAB will run the same bit of code multiple times with an iterative parameter that can change certain things about the code. If there is an element of the program that needs to carry out a process several times in a row, particularly using the previous step to compute the one after it, a for loop might be the best structure to use. A sample for loop has the following form:

```
for counter = 1:1:10
    % CODE HERE
end
```

In this line, `counter` is the variable that is getting incremented over the list. The rest of that line says that counter starts at 1, increments by 1 each loop, and stops after 10. A line of the form `counter = 2:5:34` will start at 2, increment by 5 each loop, and stop once the counter gets above 34, so after the iteration when `counter = 32`.

In order to loop through an array of values, it is useful to figure out the size of the array and use that to determine how many times the loop should be run. This sort of programming will allow your code to work for a variety of different inputs, no matter the size. This can be done with code like this.

```
v = [1,2,3,4,5]; % This will be your list of values
for counter = 1:1:length(v)
    x = v(counter)^2
end
```

To find how many elements are in an array, the `length` function will work for a linear array. If the array is more complicated, the `size` function can be used. This will give a list of values saying how large the array is in each dimension.

MATLAB also has `while` loops, which allow a loop to run up until a condition becomes false. This is better than for loops in specific situations, but either one can be used. For the code developed here, for loops will be just as easy to write as while loops.

A.5.2 If Statements

If statements, or conditional statements, allow certain parts of code to be executed only if a certain condition is met. For instance, something like

```
if counter < 5
    % CODE HERE
end
```

will only execute if the counter is less than 5, and

```
if mod(counter,2) == 0
    % CODE HERE
end
```

will only run if counter is even, that is, if the remainder when dividing counter by 2 is zero. Notice that `==` is used for comparison here to check if two things are equal, while `=` is used for variable assignment. The condition part of an if statement can be anything that gives back a true or false result. For math operations, these can be any inequalities (\leq , $<$, \geq , $>$) or `==` for testing inequality. The operator `~` is used for “not”, in that $a \sim = b$ will be true if a is not equal to b , and false if they are the same. Outside of numbers, there are other MATLAB methods that will give true or false answers. These can be things like comparing strings, but this is beyond the code developed here.

A.6 Plotting in MATLAB

Graphing in MATLAB always involves plotting a set of points, but these can be fairly easily generated from functions as well. For example

```
xPts = [1,2,3,4,5];
fx = @(x) x.^2 + 2;
yPts = [2,3,2,3,1];
figure(1);
plot(xPts, yPts);
figure(2);
plot(xPts, fx(xPts));
```

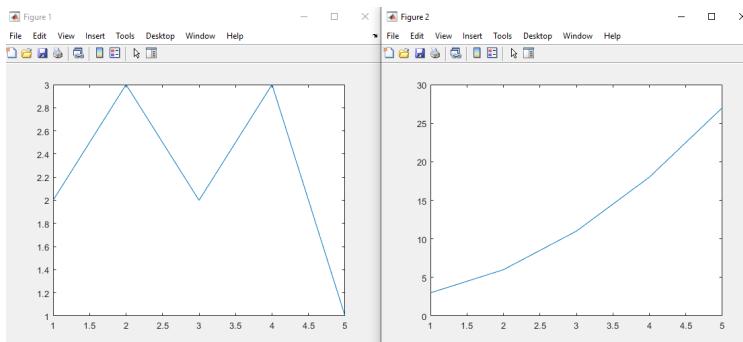


Figure A.4: Output from MATLAB plotting two graphs.

will generate two figures, referred to by the lines `figure(1)` and `figure(2)`, and allow the two graphs to be simultaneously drawn without overlapping each other. Any time MATLAB draws a plot (with the `plot` command) it will overwrite any plot that is already on the target figure. In order to put multiple plots on the same figure, the `hold on;` and `hold off;` commands can be used.

```
xPts = linspace(1,5,100);
fx = @(x) x.^2 + 2;
gx = @(x) x.^2 - 3*x + 7;
figure(1);
hold on;
plot(xPts, fx(xPts));
plot(xPts, gx(xPts));
hold off;
```

The `linspace` generates a list of 100 equally spaced values between 1 and 5 for plotting purposes. It gives an easy way to generate a lot of input values for plotting a smooth-looking graph. It also emphasizes the need to use the element-wise operations in these functions to make sure they all compute correctly.

There are many additional options that can be passed to the `plot` method in order to change the color, shape, and size of the plot. For these options, refer to the MATLAB documentation on the `plot` function at <https://www.mathworks.com/help/matlab/ref/plot.html>.

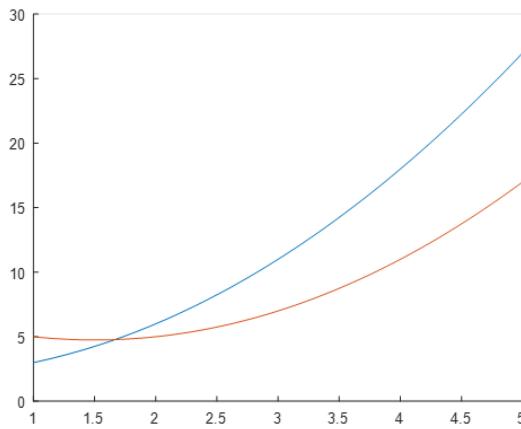


Figure A.5: Output from MATLAB plotting two functions on the same axes.

A.7 Supplemental Code Files

There are eleven supplemental code files provided. In order to use these files in a script or a Live Script, they must be placed in the same folder as the script file, so that the Current Folder window contains both the file being executed and all of these function files. Another option would be to store all of these function files in a single folder, navigating to that folder in the MATLAB Current Folder window, right-clicking on the folder, and selecting “Add to Path.” The first of these is more recommended, but the second can also work if there is a common repository to store all of the users custom MATLAB functions. The function headers are given below along with a brief description of their use.

```
function quiver244(f, t_min, t_max, y_min, y_max, col)
% quiver244.m
% Author: Matt Charnley
%
% This function draws a quiver plot for the ODE dy/dt = f(t,y) for
% t_min <= t <= t_max and y_min <= y <= y_max. The function f should be
% passed in as an anonymous function, of two variables or as a function
% handle
%
% The function draws this quiver plot in color col and saves it on the
% current figure, and generates a normalized version
% (all vectors are the same length) as the next figure,
% so that it can be accessed outside of this function.
% For this second figure, the magnitude of the arrows does not mean
% anything, but it is easier to see the direction of them.
% so that it can be accessed outside of this function. It will start with
% hold on; and end with hold off;, so the figure needs to be cleared in the
% main file if needed.
```

The main point of this function is to simplify the process of drawing quiver plots. The code here takes care of the difficulties that arise from the built-in `quiver` function in MATLAB and allows the user to input the right-hand side of a first order ODE and generate quiver plots. It will draw a quiver plot in the first figure, and a normalized quiver plot (all vectors the same length) in the second figure. It can sometimes be easier to see the general trajectory of solutions from the normalized figure, so both graphs are provided. All of the plotting commands use the `hold` commands so that they will not overwrite anything on the desired figures. This allows the overlaying of multiple plots, but means that the code calling this method must clear the figure if it needs to be cleared.

This code can be used as

```
f = @(t,y) t - exp(y);
quiver244(f, 0, 5, -6, 6, 'b');
```

```
quiver244(@f2, 0, 5, -6, 6, 'b');

function z = f2(t,y)
    z = t - exp(y);
end
```

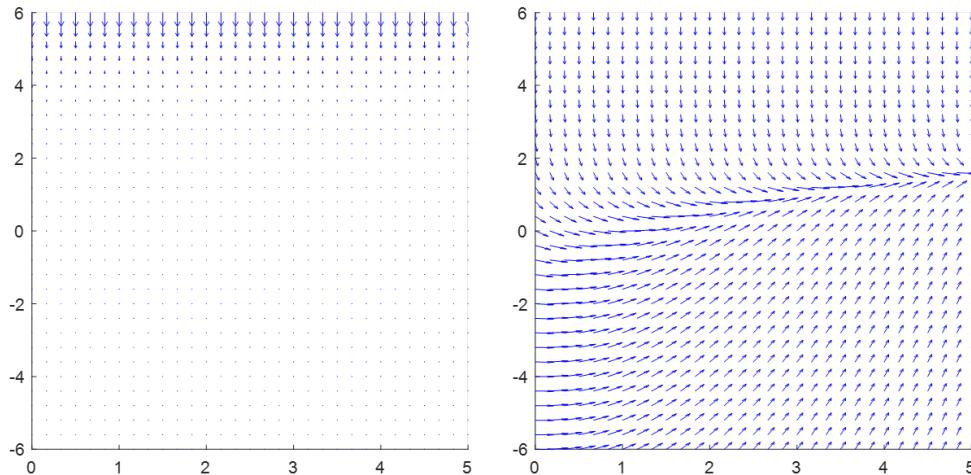


Figure A.6: Sample output from the `quiver244` function.

In each case, the ‘`b`’ indicates that the quiver plot will be drawn in blue, and the `1` before that indicates that the two plots will be drawn on figures 1 and 2.

```

function samplePlots244(f, t_min, t_max, y_min, y_max, t_0, y_0, col)
% This function takes the ODE dy/dt = f(t,y) and plots sample solutions
% with initial value (t_0, y_0). It uses ode45 to sketch out the solutions.
% t_0 must be between t_min and t_max. It also truncates the function f so
% that functions will not go off to infinity, causing this to work properly
% on vector inputs for initial conditions in y. The input y_0 can be a
% vector
% of initial values, and this function will plot a curve
% for each of those values. If using a vector of initial
% conditions, the function must be written with vector element-wise
% operations.

```

This function follows the same setup as `quiver244`, but draws sample trajectories of the solution instead of the quiver plot. It will take initial conditions as (t_0, y_0) . For a single t_0 , a vector of initial y_0 values can be passed in and the function will work correctly. This function can be used as

```

f = @(t,y) y.*(y-5).* (y+6);
samplePlots244(f, -1, 6, -7, 6, 0, [-1,0.5,4,5], 'r')

```

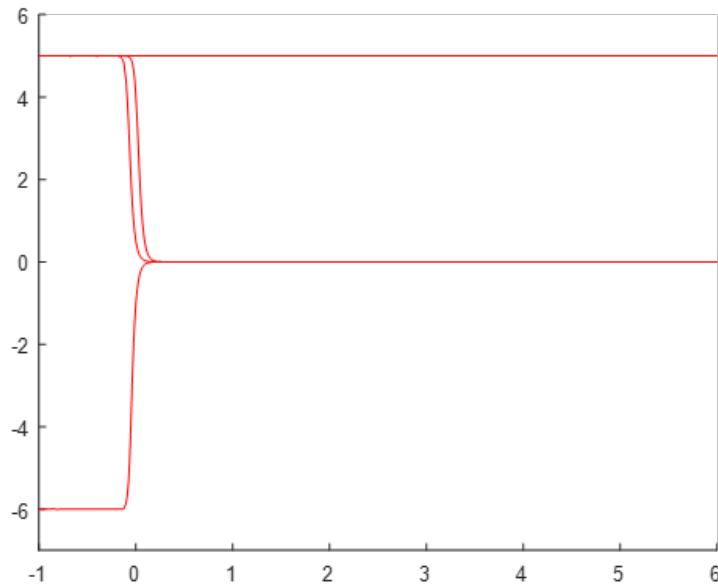


Figure A.7: Sample output from the `samplePlots244` function.

The ‘r’ here indicates that this plot will be drawn in red and put on figure 2. If this is combined with the `quiver244` method, then it will overlay these red curves on top of the quiver plot drawn on figure 2.

```

function bifDiag244(f, a_min, a_max, y_min, y_max)
% This function draws a bifurcation diagram for the ode dy/dt = f(alpha, y)
% with parameter alpha running from a_min to a_max. The axes are
% constrained to be from a_min to a_max in the horizontal direction and
% y_min to y_max in the vertical direction.
%
% The black marks are for equilibrium solutions, the blue regions are where
% the solution will tend upwards, and the red region is where it will tend
% downwards.

```

This function will draw a bifurcation diagram for the given differential equation. **Note:** This function will need the optimization tool-box add-on for MATLAB in order to run correctly. As with the previous methods, it will not overwrite the figure. Example implementation:

```

f = @(a,y) y.^2 - a.^2;
bifDiag244(f, -3, 3, -5, 5);

```

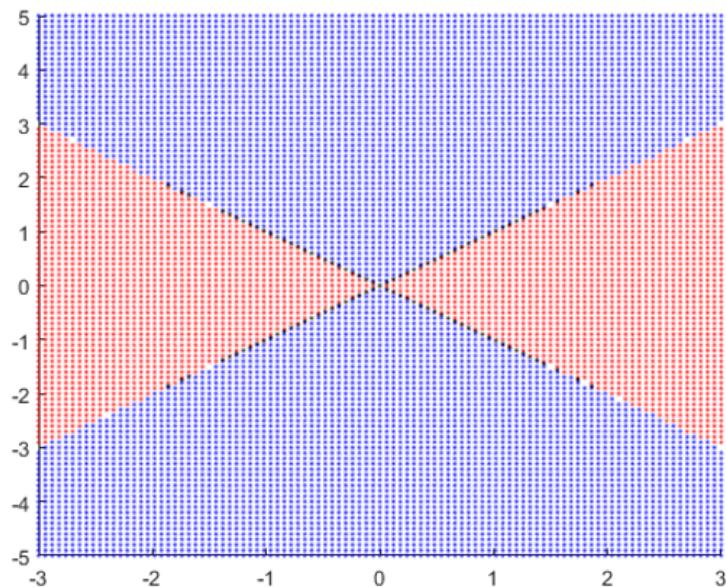


Figure A.8: Sample output from the `bifDiag244` function.

```

function quiver2D244(f,g, x_min, x_max, y_min, y_max, col)
% quiver2D244.m
% Author: Matt Charnley
%
% This function draws a quiver plot for the ODE  $dx/dt = f(x,y)$ ,  $dy/dt = g(x,y)$  for
%  $x_{\min} \leq x \leq x_{\max}$  and  $y_{\min} \leq y \leq y_{\max}$ . The functions  $f$  and  $g$  should
% be
% passed in as an anonymous functions,  $f = @(x,y) \dots$ 
%
% The function draws this quiver plot in color  $col$  in the current figure
% and generates a normalized version (all vectors are the same length)
% as the next figure, so that it can be accessed outside of this function.
% For this second figure, the magnitude of the arrows does not mean
% anything, but it is easier to see the direction of them.
%
% It will start with
% hold on; and end with hold off;, so the figure needs to be cleared in the
% main file if needed.

```

This function does the same concept as `quiver244` but for the autonomous system of differential equations

$$\frac{dx}{dt} = f(x, y) \quad \frac{dy}{dt} = g(x, y).$$

Example implementation:

```

f = @(x,y) 3.*x - 2.*x.*y;
g = @(x,y) 2.*y - 3.*x.*y;
quiver2D244(f,g, 0, 5, 0, 5, 'g');

```

```

function phaseLine(f, ymin, ymax)
% This function draws a representation of the phaseline for the
% differential equation  $dy/dt = f(y)$ . The graph is drawn from  $y_{\min}$  to  $y_{\max}$ ,
% and looks for solutions to  $f(y) = 0$  in that region to find equilibrium
% solutions. This requires the Optimization Toolbox fsolve to run
% correctly.

```

This function draws a representation of the phase line for an autonomous first order differential equation $\frac{dy}{dt} = f(y)$ from y_{\min} to y_{\max} . Example implementation:

```

f = @(y) y.*(y-3).*(y+2);
phaseLine(f, -4, 5);

```

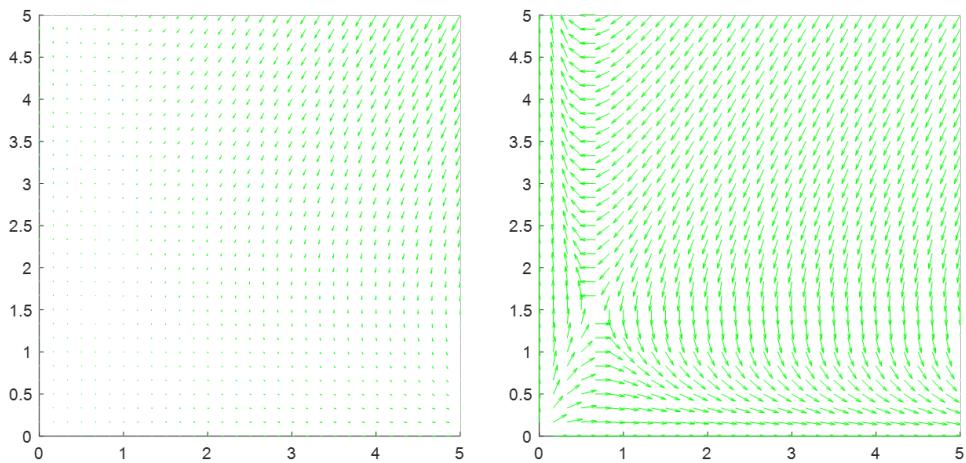


Figure A.9: Sample output from the `quiver2D244` function.

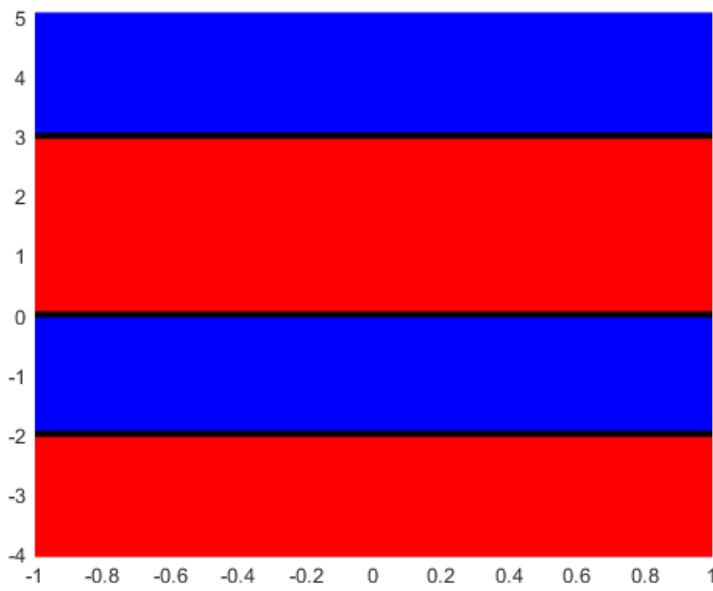


Figure A.10: Sample output from the `phaseLine` function.

```
function phasePortrait244(F, G, xmin, xmax, ymin, ymax, tmin, tmax, x0, y0)
% This function draws a 2 dimensional phase portrait for the system  $dx/dt = F(x,y)$  and  $dy/dt = G(x,y)$ . The phase portrait will be drawn with  $x$  bounds  $xmin \leq x \leq xmax$  and  $ymin \leq y \leq ymax$ . It is assumed that the initial conditions  $x0$  and  $y0$  are at  $t=0$ , with  $tmin \leq 0$  and  $tmax \geq 0$ .  $x0$  and  $y0$  can be inputted as vectors that are the same length, and a sample curve will be drawn for each of them. The black dot will always be plotted at  $\rightarrow tmin$ .
```

This function draws a phase portrait for the two-component autonomous system $\frac{dx}{dt} = F(x, y)$ and $\frac{dy}{dt} = G(x, y)$. The axes are fixed at $x_{min} \leq x \leq x_{max}$ and $y_{min} \leq y \leq y_{max}$. Solution curves are drawn starting at the (potential list of) points x_0 and y_0 , and will assume these happen at $t = 0$. The curves are drawn from t_{min} to t_{max} , and there will be a black dot plotted at t_{min} to indicate the direction of flow. Example implementation:

```
f = @(x,y) 2.*x - 3.* y;
g = @(x,y) -3.*x + y;
phasePortrait244(f, g, -3, 3, -3, 3, -2, 2, [1, 0, -1, 1, 0, -1],
                  [1,1,1,-1,-1,-1]);
```

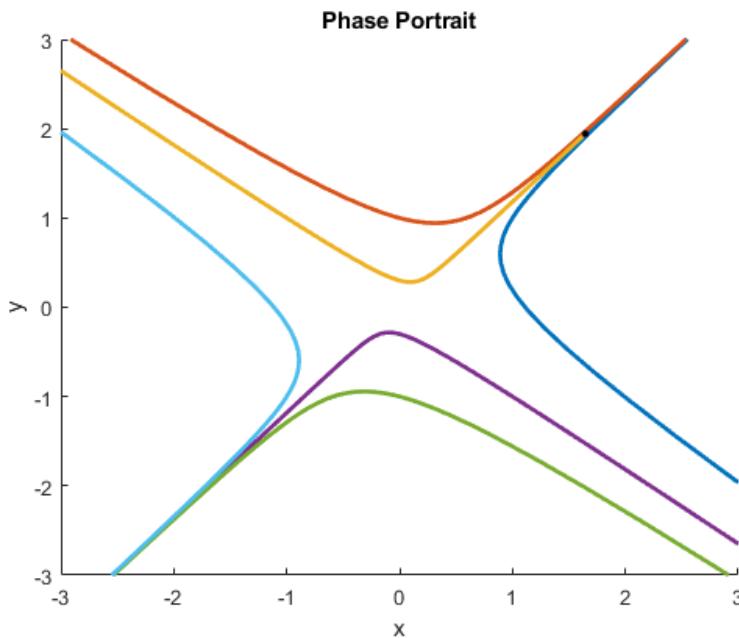


Figure A.11: Sample output from the *phasePortrait* function.

```
function [t, y] = rungeKuttaMethod(f, dt, Tf, T0, y0)
% This method solves the ODE dy/dt = f(t, y) using the Runge Kutta method
% from t=T0 to t = Tf with time step dt and initial condition y0 at t = T0.
% In this case, f should be a function of two variables, t
% (time) and y.
```

```

function [t,y] = rungeKuttaSystemMethod(f, T0, Tf, dt, y0)
% This method solves the ODE system  $dy/dt = f(t, y)$  using the Runge Kutta
% method
% from  $t=T0$  to  $t = Tf$  with time step  $dt$  and initial condition  $y0$  at  $t = T0$ .
% In this case,  $f$  should be a vector valued function of two variables,  $t$ 
% (time) and  $y$  ( $n$ -dimensional vector of unknowns). The length of the vector
%  $y0$  will determine the size of the system.

```

These two methods use the Runge-Kutta method to numerically solve the differential equation $\frac{dy}{dt} = f(t, y)$ or the system $\frac{d\vec{x}}{dt} = F(t, \vec{x})$. It will return the list of t and y values that are generated by this method.

```

function [S,I,R] = SIRModel_244(r, c, ICs, Tf)
% This code runs an SIR model for disease spread. The system of differential
% equations used here is
%  $S' = -r*S*I$ 
%  $I' = r*S*I - cI$ 
%  $R' = c*I$ 
%
% The solution is computed using the RungeKutta method, with the helper
% method rungeKuttaSystemMethod. The system is solved from  $t=0$  to  $t=Tf$ ,
% with initial conditions  $ICs$  given as a 3 component vector.

```

```

function [S,I,Q,R,D] = SIRQModel_244(alpha, beta, gamma, delta, eta, rho,
    ICs, Tf)
% This code runs a more complicated SIR model that adds in  $Q$  (a quarantined
% population) and  $D$  (a deceased population). The system of differential
% equations used here is
%  $S' = -alpha*S*I$ 
%  $I' = alpha*S*I - (beta+gamma+delta)*I$ 
%  $Q' = beta*I - (eta + rho)*Q$ 
%  $R' = gamma*I + eta*Q$ 
%  $D' = delta*I + rho*Q$ 
%
% The solution is computed using the RungeKutta method, with the helper
% method rungeKuttaSystemMethod. The system is solved from  $t=0$  to  $t=Tf$ ,
% with initial conditions  $ICs$  given as a 5 component vector.

```

```

function [S,I,Q,R,D] = SIRQVModel_244(alpha, beta, gamma, delta, eta, rho,
→ zeta, ICs, Tf)
% This code runs a more complicated SIR model that adds in Q (a quarantined
% population) and D (a deceased population). The V component adds
% vaccination into the picture, where members are moved from S to R
% directly. The system of differential equations used here is
% S' = -alpha*S*I - zeta*S
% I' = alpha*S*I - (beta+gamma+delta)*I
% Q' = beta*I - (eta + rho)*Q
% R' = gamma*I + eta*Q+zeta*S
% D' = delta*I + rho*Q
%
% The solution is computed using the RungeKutta method, with the helper
% method rungeKuttaSystemMethod. The system is solved from t=0 to t=Tf,
% with initial conditions ICs given as a 5 component vector.

```

Each of these last three methods use the Runge Kutta method to numerical solve a disease modeling problem with their respective equations. The shared arguments are the initial conditions, which are a three or five component vector depending on the problem type, and the final time T_f . The step-size used is one day, and the method will return the list of time-stepped values for each population (every day) from $t = 0$ to $t = T_f$. For SIR , the equations are

$$\frac{dS}{dt} = -rSI \quad \frac{dI}{dt} = rSI - cI \quad \frac{dR}{dt} = cI.$$

For SIRQ, the equations are

$$\begin{aligned}
\frac{dS}{dt} &= -\alpha SI \\
\frac{dI}{dt} &= \alpha SI - \beta I - \gamma I - \delta I \\
\frac{dQ}{dt} &= \beta I - \eta Q - \rho Q \\
\frac{dR}{dt} &= \gamma I + \eta Q \\
\frac{dD}{dt} &= \delta I + \rho Q
\end{aligned}$$

and for SIRQV, it is

$$\begin{aligned}\frac{dS}{dt} &= -\alpha SI - \zeta S \\ \frac{dI}{dt} &= \alpha SI - \beta I - \gamma I - \delta I \\ \frac{dQ}{dt} &= \beta I - \eta Q - \rho Q \\ \frac{dR}{dt} &= \gamma I + \eta Q + \zeta S \\ \frac{dD}{dt} &= \delta I + \rho Q\end{aligned}$$

An example implementation is

```
[S,I,R] = SIRModel_244(0.1, 0.2, [0.99; 0.01; 0], 400);
[S,I,Q,R,D] = SIRQModel_244(0.15, 0.08, 0.02, 0.03, 0.01, 0.04, [0.95; 0.05;
    ↪ 0; 0; 0], 400);
[S,I,Q,R,D] = SIRQVModel_244(0.15, 0.08, 0.02, 0.03, 0.01, 0.04, 0.2, [0.95;
    ↪ 0.05; 0; 0; 0], 400);
```


Appendix B

Prerequisite Material

This chapter provides a review of some of the material from previous classes that may be a little rusty by the time one reaches differential equations. This can be used as a reference whenever these topics come up throughout the book. A lot of this material (or inspiration for it) is taken from the Precalculus book by Stitz and Zeager [SZ].

B.1 Polynomials and Factoring

Note: Attribution: [SZ], §A.8, A.9, 2.2-2.4

There are several components of differential equations, particularly higher order equations and systems, that involve dealing with and finding roots of polynomials, using these results to generate solutions to differential equations. This appendix will review some properties of and techniques related to polynomials.

B.1.1 Definitions and Operations

First we start with the definition of a polynomial. A *polynomial* is a sum of terms each of which is a real number or a real number multiplied by one or more variables to natural number powers. Some examples of polynomials are $x^2 + x\sqrt{3} + 4$, $27x^2y + \frac{7x}{2}$ and 6. Things like $3\sqrt{x}$, $4x - \frac{2}{x+1}$ and $13x^{2/3}y^2$ are *not* polynomials. Below, we review some terminology about polynomials.

Definition B.1.1

- Terms in polynomials without variables are called *constant* terms.
- In non-constant terms, the real number factor in the expression is called the *coefficient* of the term.
- The *degree* of a non-constant term is the sum of the exponents on the variables in the term; non-zero constant terms are defined to have degree 0. The degree of a polynomial is the highest degree of the nonzero terms.
- Terms in a polynomial are called *like* terms if they have the same variables each with the same corresponding exponents.
- A polynomial is said to be *simplified* if all arithmetic operations have been completed and there are no longer any like terms.
- A simplified polynomial is called a
 - *monomial* if it has exactly one nonzero term
 - *binomial* if it has exactly two nonzero terms
 - *trinomial* if it has exactly three nonzero terms

For example, $x^2 + x\sqrt{3} + 4$ is a trinomial of degree 2. The coefficient of x^2 is 1 and the constant term is 4. The polynomial $27x^2y + \frac{7x}{2}$ is a binomial of degree 3 ($x^2y = x^2y^1$) with constant term 0.

The concept of ‘like’ terms really amounts to finding terms which can be combined using the Distributive Property. For example, in the polynomial $17x^2y - 3xy^2 + 7xy^2$, $-3xy^2$ and $7xy^2$ are like terms, since they have the same variables with the same corresponding

exponents. This allows us to combine these two terms as follows:

$$17x^2y - 3xy^2 + 7xy^2 = 17x^2y + (-3)xy^2 + 7xy^2 + 17x^2y + (-3 + 7)xy^2 = 17x^2y + 4xy^2$$

Note that even though $17x^2y$ and $4xy^2$ have the same variables, they are not like terms since in the first term we have x^2 and $y = y^1$ but in the second we have $x = x^1$ and $y = y^2$ so the corresponding exponents aren't the same. Hence, $17x^2y + 4xy^2$ is the simplified form of the polynomial.

There are four basic operations we can perform with polynomials: addition, subtraction, multiplication and division. Addition, subtraction, and multiplication follow the standard properties of real numbers after distributing or expanding all terms (for multiplication) and then collecting like terms again. Division, on the other hand, is a bit more complicated and will be discussed next.

Polynomial Long Division

We now turn our attention to polynomial long division. Dividing two polynomials follows the same algorithm, in principle, as dividing two natural numbers so we review that process first. Suppose we wished to divide 2585 by 79. The standard division tableau is given below.

$$\begin{array}{r} 32 \\ 79 \overline{)2585} \\ -237 \downarrow \\ \hline 215 \\ -158 \\ \hline 57 \end{array}$$

In this case, 79 is called the *divisor*, 2585 is called the *dividend*, 32 is called the *quotient* and 57 is called the *remainder*. We can check our answer by showing:

$$\text{dividend} = (\text{divisor})(\text{quotient}) + \text{remainder}$$

or in this case, $2585 = (79)(32) + 57\checkmark$. We hope that the long division tableau evokes warm, fuzzy memories of your formative years as opposed to feelings of hopelessness and frustration. If you experience the latter, keep in mind that the Division Algorithm essentially is a two-step process, iterated over and over again. First, we guess the number of times the divisor goes into the dividend and then we subtract off our guess. We repeat those steps with what's left over until what's left over (the remainder) is less than what we started with (the divisor). That's all there is to it!

The division algorithm for polynomials has the same basic two steps but when we subtract polynomials, we must take care to subtract *like terms* only. As a transition to polynomial division, let's write out our previous division tableau in expanded form.

$$\begin{array}{r}
 & 3 \cdot 10 + 2 \\
 7 \cdot 10 + 9 & \overline{)2 \cdot 10^3 + 5 \cdot 10^2 + 8 \cdot 10 + 5} \\
 & - (2 \cdot 10^3 + 3 \cdot 10^2 + 7 \cdot 10) \quad \downarrow \\
 & 2 \cdot 10^2 + 1 \cdot 10 + 5 \\
 & - (1 \cdot 10^2 + 5 \cdot 10 + 8) \\
 & 5 \cdot 10 + 7
 \end{array}$$

Written this way, we see that when we line up the digits we are really lining up the coefficients of the corresponding powers of 10 - much like how we'll have to keep the powers of x lined up in the same columns. The big difference between polynomial division and the division of natural numbers is that the value of x is an unknown quantity. So unlike using the known value of 10, when we subtract there can be no regrouping of coefficients as in our previous example. (The subtraction $215 - 158$ requires us to ‘regroup’ or ‘borrow’ from the tens digit, then the hundreds digit.) This actually makes polynomial division easier.* Before we dive into examples, we first state a theorem telling us when we can divide two polynomials, and what to expect when we do so.

Theorem B.1.1 (Polynomial Division)

Let d and p be nonzero polynomials where the degree of p is greater than or equal to the degree of d . There exist two unique polynomials, q and r , such that $p = d \cdot q + r$, where either $r = 0$ or the degree of r is strictly less than the degree of d .

Essentially, Theorem B.1.1 tells us that we can divide polynomials whenever the degree of the divisor is less than or equal to the degree of the dividend. We know we're done with the division when the polynomial left over (the remainder) has a degree strictly less than the divisor. It's time to walk through a few examples to refresh your memory.

Example B.1.1: Perform the indicated division. Check your answer by showing

$$\text{dividend} = (\text{divisor})(\text{quotient}) + \text{remainder}$$

- | | |
|--|------------------------------------|
| 1. $(x^3 + 4x^2 - 5x - 14) \div (x - 2)$ | 2. $(2t + 7) \div (3t - 4)$ |
| 3. $(6y^2 - 1) \div (2y + 5)$ | 4. $(w^3) \div (w^2 - \sqrt{2})$. |

Solution:

- To begin $(x^3 + 4x^2 - 5x - 14) \div (x - 2)$, we divide the first term in the dividend, namely x^3 , by the first term in the divisor, namely x , and get $\frac{x^3}{x} = x^2$. This then becomes the first term in the quotient. We proceed as in regular long division at this point: we multiply the entire divisor, $x - 2$, by this first term in the quotient to get

*In our opinion - you can judge for yourself.

$x^2(x - 2) = x^3 - 2x^2$. We then subtract this result from the dividend.

$$\begin{array}{r} x^2 \\ x-2 \overline{)x^3 + 4x^2 - 5x - 14} \\ \underline{-(x^3 - 2x^2)} \quad \downarrow \\ 6x^2 - 5x \end{array}$$

Now we ‘bring down’ the next term of the quotient, namely $-5x$, and repeat the process. We divide $\frac{6x^2}{x} = 6x$, and add this to the quotient polynomial, multiply it by the divisor (which yields $6x(x - 2) = 6x^2 - 12x$) and subtract.

$$\begin{array}{r} x^2 + 6x \\ x-2 \overline{)x^3 + 4x^2 - 5x - 14} \\ \underline{-(x^3 - 2x^2)} \quad \downarrow \\ 6x^2 - 5x \quad \downarrow \\ \underline{-(6x^2 - 12x)} \quad \downarrow \\ 7x - 14 \end{array}$$

Finally, we ‘bring down’ the last term of the dividend, namely -14 , and repeat the process. We divide $\frac{7x}{x} = 7$, add this to the quotient, multiply it by the divisor (which yields $7(x - 2) = 7x - 14$) and subtract.

$$\begin{array}{r} x^2 + 6x + 7 \\ x-2 \overline{)x^3 + 4x^2 - 5x - 14} \\ \underline{-(x^3 - 2x^2)} \\ 6x^2 - 5x \\ \underline{-(6x^2 - 12x)} \\ 7x - 14 \\ \underline{-(7x - 14)} \\ 0 \end{array}$$

In this case, we get a quotient of $x^2 + 6x + 7$ with a remainder of 0. To check our answer, we compute

$$(x - 2)(x^2 + 6x + 7) + 0 = x^3 + 6x^2 + 7x - 2x^2 - 12x - 14 = x^3 + 4x^2 - 5x - 14 \checkmark$$

2. To compute $(2t + 7) \div (3t - 4)$, we start as before. We find $\frac{2t}{3t} = \frac{2}{3}$, so that becomes the first (and only) term in the quotient. We multiply the divisor $(3t - 4)$ by $\frac{2}{3}$ and get

$2t - \frac{8}{3}$. We subtract this from the divided and get $\frac{29}{3}$.

$$\begin{array}{r} & \frac{2}{3} \\ & \frac{3}{3} \\ 3t-4 & \overline{)2t + 7} \\ -\left(2t - \frac{8}{3}\right) \\ \hline & \frac{29}{3} \end{array}$$

Our answer is $\frac{2}{3}$ with a remainder of $\frac{29}{3}$. To check our answer, we compute

$$(3t - 4) \left(\frac{2}{3}\right) + \frac{29}{3} = 2t - \frac{8}{3} + \frac{29}{3} = 2t + \frac{21}{3} = 2t + 7 \checkmark$$

3. When we set-up the tableau for $(6y^2 - 1) \div (2y + 5)$, we must first issue a ‘placeholder’ for the ‘missing’ y -term in the dividend, $6y^2 - 1 = 6y^2 + 0y - 1$. We then proceed as before. Since $\frac{6y^2}{2y} = 3y$, $3y$ is the first term in our quotient. We multiply $(2y + 5)$ times $3y$ and subtract it from the dividend. We bring down the -1 , and repeat.

$$\begin{array}{r} 3y \quad - \frac{15}{2} \\ \hline 2y+5 \overline{)6y^2 + 0y - 1} \\ -(6y^2 + 15y) \quad \downarrow \\ \hline -15y - 1 \\ -\left(-15y - \frac{75}{2}\right) \\ \hline \frac{73}{2} \end{array}$$

Our answer is $3y - \frac{15}{2}$ with a remainder of $\frac{73}{2}$. To check our answer, we compute:

$$(2y + 5) \left(3y - \frac{15}{2}\right) + \frac{73}{2} = 6y^2 - 15y + 15y - \frac{75}{2} + \frac{73}{2} = 6y^2 - 1 \checkmark$$

4. For our last example, we need ‘placeholders’ for both the divisor $w^2 - \sqrt{2} = w^2 + 0w - \sqrt{2}$ and the dividend $w^3 = w^3 + 0w^2 + 0w + 0$. The first term in the quotient is $\frac{w^3}{w^2} = w$, and when we multiply and subtract this from the dividend, we’re left with just $0w^2 + w\sqrt{2} + 0 = w\sqrt{2}$.

$$\begin{array}{r} w \\ \hline w^2 + 0w - \sqrt{2} \overline{)w^3 + 0w^2 + 0w + 0} \\ - \left(w^3 + 0w^2 - w\sqrt{2}\right) \quad \downarrow \\ \hline 0w^2 + w\sqrt{2} + 0 \end{array}$$

Since the degree of $w\sqrt{2}$ (which is 1) is less than the degree of the divisor (which is 2), we are done.* Our answer is w with a remainder of $w\sqrt{2}$. To check, we compute:

$$(w^2 - \sqrt{2})w + w\sqrt{2} = w^3 - w\sqrt{2} + w\sqrt{2} = w^3 \checkmark$$

□

B.1.2 Synthetic Division

Usually, when we want to divide polynomials, it is because we are trying to find all roots of a polynomial. This comes from the idea that if we have a polynomial $p(x)$ and a value x_0 so that $p(x_0) = 0$, then x_0 is a root of the polynomial. This means that $(x - x_0)$ is a factor of $p(x)$, so that we can write

$$p(x) = (x - x_0)q(x)$$

where $q(x)$ is a polynomial with one lower degree than p . We can find this $q(x)$ by dividing

$$q(x) = \frac{p(x)}{x - x_0},$$

which is why we need division to sort this out.

This means that we need to find the roots (or at least a root) to know what to divide $p(x)$ by in order to start this process. The main theorem that can tell us where to start is the Rational Roots Theorem.

Theorem B.1.2 (Rational Zeros Theorem)

Suppose $f(x) = a_nx^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$ is a polynomial of degree n with $n \geq 1$, and a_0, a_1, \dots, a_n are integers. If r is a rational zero of f , then r is of the form $\pm \frac{p}{q}$, where p is a factor of the constant term a_0 , and q is a factor of the leading coefficient a_n .

The Rational Zeros Theorem gives us a list of numbers to try in our synthetic division and that is a lot nicer than simply guessing. If none of the numbers in the list are zeros, then either the polynomial has no real zeros at all, or all of the real zeros are irrational numbers.

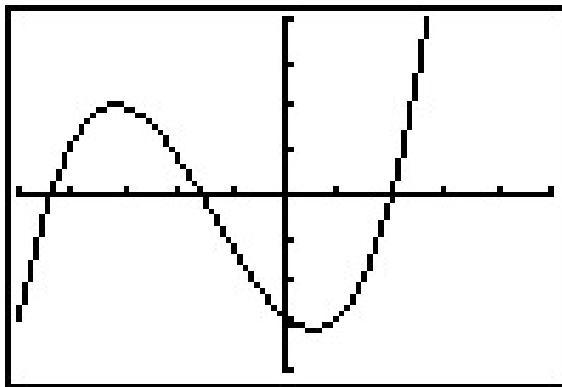
Example B.1.2: Let $f(x) = 2x^4 + 4x^3 - x^2 - 6x - 3$. Use the Rational Zeros Theorem to list all of the possible rational zeros of f .

Solution: To generate a complete list of rational zeros, we need to take each of the factors of constant term, $a_0 = -3$, and divide them by each of the factors of the leading coefficient $a_4 = 2$. The factors of -3 are ± 1 and ± 3 . Since the Rational Zeros Theorem tacks on a \pm anyway, for the moment, we consider only the positive factors 1 and 3. The factors of 2 are 1 and 2, so the Rational Zeros Theorem gives the list $\{\pm \frac{1}{1}, \pm \frac{1}{2}, \pm \frac{3}{1}, \pm \frac{3}{2}\}$ or $\{\pm \frac{1}{2}, \pm 1, \pm \frac{3}{2}, \pm 3\}$. □

*Since $\frac{0w^2}{w^2} = 0$, we could proceed, write our quotient as $w + 0$, and move on... but even pedants have limits.

But this still doesn't make the process easy or straight-forward for finding the roots. How can we take this list of options and easily figure out where the roots are, and what the remaining polynomial $q(x)$ is?

We start by way of example: suppose we wish to determine the zeros of $f(x) = x^3 + 4x^2 - 5x - 14$. Setting $f(x) = 0$ results in the polynomial equation $x^3 + 4x^2 - 5x - 14 = 0$. Despite all of the factoring techniques we learned (and forgot!), this equation foils* us at every turn. Knowing that the zeros of f correspond to x -intercepts on the graph of $y = f(x)$, we use a graphing utility to produce the graph below on the left. The graph suggests that the function has three zeros, one of which appears to be $x = 2$ and two others for whom we are provided what we assume to be decimal approximations: $x \approx -4.414$ and $x \approx -1.586$. We can verify if these are zeros easily enough. We find $f(2) = (2)^3 + 4(2)^2 - 5(2) - 14 = 0$, but $f(-4.414) \approx 0.0039$ and $f(-1.586) \approx 0.0022$. While these last two values are probably by some measures, 'close' to 0, they are not *exactly* equal to 0. The question becomes: is there a way to use the fact that $x = 2$ is a zero to obtain the other two zeros? Based on our experience, if $x = 2$ is a zero, it seems that there should be a factor of $(x - 2)$ lurking around in the factorization of $f(x)$. In other words, we should expect that $x^3 + 4x^2 - 5x - 14 = (x - 2)q(x)$, where $q(x)$ is some other polynomial. How could we find such a $q(x)$, if it even exists? The answer comes from our old friend, polynomial division. Below on the right, we perform the long division: $(x^3 + 4x^2 - 5x - 14) \div (x - 2)$ and obtain $x^2 + 6x + 7$.



$$\begin{array}{r}
 x^2 + 6x + 7 \\
 x-2 \overline{)x^3 + 4x^2 - 5x - 14} \\
 -(x^3 - 2x^2) \\
 \hline
 6x^2 - 5x \\
 -(6x^2 - 12x) \\
 \hline
 7x - 14 \\
 -(7x - 14) \\
 \hline
 0
 \end{array}$$

Said differently, $f(x) = x^3 + 4x^2 - 5x - 14 = (x - 2)(x^2 + 6x + 7)$. Using this form of $f(x)$, we find the zeros by solving $(x - 2)(x^2 + 6x + 7) = 0$. Setting each factor equal to 0, we get $x - 2 = 0$ (which gives us our known zero, $x = 2$) as well as $x^2 + 6x + 7 = 0$. The latter doesn't factor nicely, so we apply the Quadratic Formula to get $x = -3 \pm \sqrt{2}$. Sure enough, $-3 - \sqrt{2} \approx -4.414$ and $-3 + \sqrt{2} \approx -1.586$. We leave it to the reader to show $f(-3 - \sqrt{2}) = 0$ and $f(-3 + \sqrt{2}) = 0$.

The point of this section is to generalize the technique applied here. First up is a friendly reminder of what we can expect when we divide polynomials.

*pun intended

Theorem B.1.3

Suppose $d(x)$ and $p(x)$ are nonzero polynomial functions where the degree of p is greater than or equal to the degree of d . There exist two unique polynomial functions, $q(x)$ and $r(x)$, such that $p(x) = d(x)q(x) + r(x)$, where either $r(x) = 0$ or the degree of r is strictly less than the degree of d .

As you may recall, all of the polynomials in Theorem B.1.3 have special names. The polynomial p is called the *dividend*; d is the *divisor*; q is the *quotient*; r is the *remainder*. If $r(x) = 0$ then d is called a *factor* of p . The word ‘unique’ here is critical in that it guarantees there is only *one* quotient and remainder for each division problem.* The proof of Theorem B.1.3 is usually relegated to a course in Abstract Algebra, but we can still use the result to move forward with the rest of this section.

If we want to find all of the roots of a polynomial in a reasonable way, we had better find a more efficient way to divide polynomial functions by quantities of the form $x - c$. Fortunately, people like Ruffini and Horner have already blazed this trail. Let’s take a closer look at the long division we performed at the beginning of the section and try to streamline it. First off, let’s change all of the subtractions into additions by distributing through the -1 s.

$$\begin{array}{r} x^2 + 6x + 7 \\ x-2 \overline{)x^3 + 4x^2 - 5x - 14} \\ -x^3 + 2x^2 \\ \hline 6x^2 - 5x \\ -6x^2 + 12x \\ \hline 7x - 14 \\ -7x + 14 \\ \hline 0 \end{array}$$

Next, observe that the terms $-x^3$, $-6x^2$ and $-7x$ are the exact opposite of the terms above them. The algorithm we use ensures this is always the case, so we can omit them without losing any information. Also note that the terms we ‘bring down’ (namely the $-5x$ and -14) aren’t really necessary to recopy, so we omit them, too.

$$\begin{array}{r} x^2 + 6x + 7 \\ x-2 \overline{)x^3 + 4x^2 - 5x - 14} \\ 2x^2 \\ \hline 6x^2 \\ 12x \\ \hline 7x \\ 14 \\ \hline 0 \end{array}$$

Let’s move terms up a bit and copy the x^3 into the last row.

*Hence the use of the definite article ‘the’ when speaking of *the* quotient and *the* remainder.

$$\begin{array}{r} x^2 + 6x + 7 \\ x-2 \overline{)x^3 + 4x^2 - 5x - 14} \\ \quad 2x^2 \quad 12x \quad 14 \\ \hline \quad x^3 \quad 6x^2 \quad 7x \quad 0 \end{array}$$

Note that by arranging things in this manner, each term in the last row is obtained by adding the two terms above it. Notice also that the quotient polynomial can be obtained by dividing each of the first three terms in the last row by x and adding the results. If you take the time to work back through the original division problem, you will find that this is exactly the way we determined the quotient polynomial. This means that we no longer need to write the quotient polynomial down, nor the x in the divisor, to determine our answer.

$$\begin{array}{r} -2 \mid x^3 + 4x^2 - 5x - 14 \\ \quad 2x^2 \quad 12x \quad 14 \\ \hline \quad x^3 \quad 6x^2 \quad 7x \quad 0 \end{array}$$

We've streamlined things quite a bit so far, but we can still do more. Let's take a moment to remind ourselves where the $2x^2$, $12x$ and 14 came from in the second row. Each of these terms was obtained by multiplying the terms in the quotient, x^2 , $6x$ and 7 , respectively, by the -2 in $x - 2$, then by -1 when we changed the subtraction to addition. Multiplying by -2 then by -1 is the same as multiplying by 2 , so we replace the -2 in the divisor by 2 . Furthermore, the coefficients of the quotient polynomial match the coefficients of the first three terms in the last row, so we now take the plunge and write only the coefficients of the terms to get

$$\begin{array}{r} 2 \mid 1 \quad 4 \quad -5 \quad -14 \\ \quad 2 \quad 12 \quad 14 \\ \hline \quad 1 \quad 6 \quad 7 \quad 0 \end{array}$$

We have constructed a *synthetic division tableau* for this polynomial division problem. Let's re-work our division problem using this tableau to see how it greatly streamlines the division process. To divide $x^3 + 4x^2 - 5x - 14$ by $x - 2$, we write 2 in the place of the divisor and the coefficients of $x^3 + 4x^2 - 5x - 14$ in for the dividend. Then 'bring down' the first coefficient of the dividend.

$$\begin{array}{r} 2 \mid 1 \quad 4 \quad -5 \quad -14 \\ \hline \end{array} \qquad \begin{array}{r} 2 \mid 1 \quad 4 \quad -5 \quad -14 \\ \downarrow \\ \hline 1 \end{array}$$

Next, take the 2 from the divisor and multiply by the 1 that was 'brought down' to get 2 . Write this underneath the 4 , then add to get 6 .

$$\begin{array}{r} 2 \mid 1 \quad 4 \quad -5 \quad -14 \\ \downarrow \quad 2 \\ \hline 1 \end{array} \qquad \begin{array}{r} 2 \mid 1 \quad 4 \quad -5 \quad -14 \\ \downarrow \quad 2 \\ \hline 1 \quad 6 \end{array}$$

Now take the 2 from the divisor times the 6 to get 12, and add it to the -5 to get 7.

$$\begin{array}{r} 2 | \begin{array}{rrrr} 1 & 4 & -5 & -14 \\ \downarrow & 2 & 12 & \\ \hline 1 & 6 & & \end{array} \end{array}$$

$$\begin{array}{r} 2 | \begin{array}{rrrr} 1 & 4 & -5 & -14 \\ \downarrow & 2 & 12 & \\ \hline 1 & 6 & 7 & \end{array} \end{array}$$

Finally, take the 2 in the divisor times the 7 to get 14, and add it to the -14 to get 0.

$$\begin{array}{r} 2 | \begin{array}{rrrr} 1 & 4 & -5 & -14 \\ \downarrow & 2 & 12 & 14 \\ \hline 1 & 6 & 7 & \end{array} \end{array}$$

$$\begin{array}{r} 2 | \begin{array}{rrrr} 1 & 4 & -5 & -14 \\ \downarrow & 2 & 12 & 14 \\ \hline 1 & 6 & 7 & [0] \end{array} \end{array}$$

The first three numbers in the last row of our tableau are the coefficients of the quotient polynomial. Remember, we started with a third degree polynomial and divided by a first degree polynomial, so the quotient is a second degree polynomial. Hence the quotient is $x^2 + 6x + 7$. The number in the box is the remainder. Synthetic division is our tool of choice for dividing polynomials by divisors of the form $x - c$. It is important to note that it works *only* for these kinds of divisors.* Also take note that when a polynomial (of degree at least 1) is divided by $x - c$, the result will be a polynomial of exactly one less degree. Finally, it is worth the time to trace each step in synthetic division back to its corresponding step in long division. While the authors have done their best to indicate where the algorithm comes from, there is no substitute for working through it yourself.

Example B.1.3: Use synthetic division to perform the following polynomial divisions. Identify the quotient and remainder.

$$\begin{array}{lll} 1. (5x^3 - 2x^2 + 1) \div (x - 3) & 2. (t^3 + 8) \div (t + 2) & 3. \frac{4 - 8z - 12z^2}{2z - 3} \end{array}$$

Solution:

- When setting up the synthetic division tableau, the coefficients of even ‘missing’ terms need to be accounted for, so we enter 0 for the coefficient of x in the dividend.

$$\begin{array}{r} 3 | \begin{array}{rrrr} 5 & -2 & 0 & 1 \\ \downarrow & 15 & 39 & 117 \\ \hline 5 & 13 & 39 & [118] \end{array} \end{array}$$

Since the dividend was a third degree polynomial function, the quotient is a second degree (quadratic) polynomial function with coefficients 5, 13 and 39: $q(x) = 5x^2 + 13x + 39$. The remainder is $r(x) = 118$. According to Theorem B.1.3, we have $5x^3 - 2x^2 + 1 = (x - 3)(5x^2 + 13x + 39) + 118$, which we leave to the reader to check.

*You’ll need to use good old-fashioned polynomial long division for divisors of degree larger than 1.

2. To use synthetic division here, we rewrite $t + 2$ as $t - (-2)$ and proceed as before

$$\begin{array}{r|rrrr} -2 & 1 & 0 & 0 & 8 \\ \downarrow & -2 & 4 & -8 \\ \hline 1 & -2 & 4 & \boxed{0} \end{array}$$

We get the quotient $q(t) = t^2 - 2t + 4$ and the remainder $r(t) = 0$. Relating the dividend, quotient and remainder gives: $t^3 + 8 = (t + 2)(t^2 - 2t + 4)$, which is a specific instance of the ‘sum of cubes’ formula some of you may recall.

3. To divide $4 - 8z - 12z^2$ by $2z - 3$, two things must be done. First, we write the dividend in descending powers of z as $-12z^2 - 8z + 4$. Second, since synthetic division works only for factors of the form $z - c$, we factor $2z - 3$ as $2(z - \frac{3}{2})$. Hence, we are dividing $-12z^2 - 8z + 4$ by two factors: 2 and $(z - \frac{3}{2})$. Dividing first by 2, we obtain $-6z^2 - 4z + 2$. Next, we divide $-6z^2 - 4z + 2$ by $(z - \frac{3}{2})$:

$$\begin{array}{r|rrr} \frac{3}{2} & -6 & -4 & 2 \\ \downarrow & -9 & -\frac{39}{2} \\ \hline -6 & -13 & \boxed{-\frac{35}{2}} \end{array}$$

Hence, $-6z^2 - 4z + 2 = (z - \frac{3}{2})(-6z - 13) - \frac{35}{2}$. However when it comes to writing the dividend, quotient and remainder in the form given in Theorem B.1.3, we need to find $q(z)$ and $r(z)$ so that $-12z^2 - 8z + 4 = (2z - 3)q(z) + r(z)$. Hence, starting with $-6z^2 - 4z + 2 = (z - \frac{3}{2})(-6z - 13) - \frac{35}{2}$, we multiply 2 back on both sides:

$$\begin{aligned} -6z^2 - 4z + 2 &= (z - \frac{3}{2})(-6z - 13) - \frac{35}{2} \\ 2(-6z^2 - 4z + 2) &= 2[(z - \frac{3}{2})(-6z - 13) - \frac{35}{2}] \\ -12z^2 - 8z + 4 &= 2(z - \frac{3}{2})(-6z - 13) - 2(\frac{35}{2}) \\ -12z^2 - 8z + 4 &= (2z - 3)(-6z - 13) - 35 \end{aligned}$$

At this stage, we have written $-12z^2 - 8z + 4$ in the form $(2z - 3)q(z) + r(z)$, so we identify the quotient as $q(z) = -6z - 13$ and the remainder is $r(z) = -35$. But how can we be sure these are the same quotient and remainder polynomial functions we would have obtained if we had taken the time to do the long division in the first place? Because of the word ‘unique’ in Theorem B.1.3. The theorem states that there is only one way to decompose $-12z^2 - 8z + 4$ as $(2z - 3)q(z) + r(z)$. Since we have found such a way, we can be sure it is the only way.*

The next example pulls together all of the concepts discussed in this section.

Example B.1.4: Let $p(x) = 2x^3 - 5x + 3$.

1. Find $p(-2)$ using The Remainder Theorem. Check your answer by substitution.

*But it wouldn’t hurt to check, just this once.

2. Verify $x = 1$ is a zero of p and use this information to all the real zeros of p .

Solution:

1. The Remainder Theorem states $p(-2)$ is the remainder when $p(x)$ is divided by $x - (-2)$.

We set up our synthetic division tableau below. We are careful to record the coefficient of x^2 as 0:

$$\begin{array}{r|rrrr} -2 & 2 & 0 & -5 & 3 \\ \downarrow & -4 & 8 & -6 & \\ \hline 2 & -4 & 3 & \boxed{-3} \end{array}$$

According to the Remainder Theorem, $p(-2) = -3$. We can check this by direct substitution into the formula for $p(x)$: $p(-2) = 2(-2)^3 - 5(-2) + 3 = -16 + 10 + 3 = -3$.

2. We verify $x = 1$ is a zero of p by evaluating $p(1) = 2(1)^3 - 5(1) + 3 = 0$. To see if there are any more real zeros, we need to solve $p(x) = 2x^3 - 5x + 3 = 0$. From the Factor Theorem, we know since $p(1) = 0$, we can factor $p(x)$ as $(x - 1)q(x)$. To find $q(x)$, we use synthetic division:

$$\begin{array}{r|rrrr} 1 & 2 & 0 & -5 & 3 \\ \downarrow & 2 & 2 & -3 & \\ \hline 2 & 2 & -3 & \boxed{0} \end{array}$$

As promised, our remainder is 0, and we get $p(x) = (x - 1)(2x^2 + 2x - 3)$. Setting this form of $p(x)$ equal to 0 we get $(x - 1)(2x^2 + 2x - 3) = 0$. We recover $x = 1$ from setting $x - 1 = 0$ but we also obtain $x = \frac{-1 \pm \sqrt{7}}{2}$ from $2x^2 + 2x - 3 = 0$, courtesy of the Quadratic Formula.

□

Our next example demonstrates how we can extend the synthetic division tableau to accommodate zeros of multiplicity greater than 1.

Example B.1.5: Let $p(x) = 4x^4 - 4x^3 - 11x^2 + 12x - 3$. Show $x = \frac{1}{2}$ is a zero of multiplicity 2 and find all of the remaining real zeros of p .

Solution: While computing $p\left(\frac{1}{2}\right) = 0$ shows $x = \frac{1}{2}$ is a zero of p , to prove it has multiplicity 2, we need to factor $p(x) = (x - \frac{1}{2})^2 q(x)$ with $q\left(\frac{1}{2}\right) \neq 0$. We set up for synthetic division, but instead of stopping after the first division, we continue the tableau downwards and divide $(x - \frac{1}{2})$ directly into the quotient we obtained from the first division as follows:

$$\begin{array}{r|rrrrr} \frac{1}{2} & 4 & -4 & -11 & 12 & -3 \\ \downarrow & 2 & -1 & -6 & 3 & \\ \hline \frac{1}{2} & 4 & -2 & -12 & 6 & \boxed{0} \\ \downarrow & 2 & 0 & -6 & & \\ \hline 4 & 0 & -12 & \boxed{0} & & \end{array}$$

We get:^{*} $4x^4 - 4x^3 - 11x^2 + 12x - 3 = (x - \frac{1}{2})^2 (4x^2 - 12)$. Note if we let $q(x) = 4x^2 - 12$, then $q(\frac{1}{2}) = 4(\frac{1}{2})^2 - 12 = -11 \neq 0$ which proves $x = \frac{1}{2}$ is a zero of p of multiplicity 2. To find the remaining zeros of p , we set the quotient $4x^2 - 12 = 0$, so $x^2 = 3$ and extract square roots to get $x = \pm\sqrt{3}$. \square

One last wrinkle in this process is complex roots, since it is possible for a polynomial (particularly a quadratic polynomial) to have complex numbers as roots. For a reminder of some more properties of complex numbers see § B.2. For this section in particular, we only need a few basic facts.

For us, it suffices to review the basic vocabulary.

Definition B.1.2

- The imaginary unit $i = \sqrt{-1}$ satisfies the two following properties
 1. $i^2 = -1$
 2. If c is a real number with $c \geq 0$ then $\sqrt{-c} = i\sqrt{c}$
- The *complex numbers* are the set of numbers $\mathbb{C} = \{a + bi \mid a, b \in \mathbb{R}\}$
- Given a complex number $z = a + bi$, the *complex conjugate* of z , $\bar{z} = \overline{a + bi} = a - bi$.

Note that every real number is a complex number, that is $\mathbb{R} \subseteq \mathbb{C}$. To see this, take your favorite real number, say 117. We may write $117 = 117 + 0i$ which puts in the form $a + bi$. Hence, when we speak of the ‘complex zeros’ of a polynomial function, we are talking about not just the non-real, but also the real zeros.

Complex numbers, by their very definition, are two dimensional creatures. To see this, we may identify a complex number $z = a + bi$ with the point in the Cartesian plane (a, b) . The horizontal axis is called the ‘real’ axis since points here have the form $(a, 0)$ which corresponds to numbers of the form $z = a + 0i = a$ which are the real numbers. The vertical axis is called the ‘imaginary’ axis since points here are of the form $(0, b)$ which correspond to numbers of the form $z = 0 + bi = bi$, the so-called ‘purely imaginary’ numbers. Below we plot some complex numbers on this so-called ‘Complex Plane.’ Plotting a set of complex numbers this way is called an [Argand Diagram](#), and opens up a wealth of opportunities to explore many algebraic properties of complex numbers geometrically. For example, complex conjugation amounts to a reflection about the real axis, and multiplication by i amounts to a 90° rotation. While we won’t have much use for the Complex Plane in this section, it is worth introducing this concept now, if, for no other reason, it gives the reader a sense of the vastness of the complex number system and the role of the real numbers in it.

Returning to zeros of polynomials, suppose we wish to find the zeros of $f(x) = x^2 - 2x + 5$. To solve the equation $x^2 - 2x + 5 = 0$, we note that the quadratic doesn’t factor nicely, so we

^{*}For those wanting more detail: the first division gives: $4x^4 - 4x^3 - 11x^2 + 12x - 3 = (x - \frac{1}{2})(4x^3 - 2x^2 - 12x + 6)$. The second division gives: $4x^3 - 2x^2 - 12x + 6 = (x - \frac{1}{2})(4x^2 - 12)$.

resort to the Quadratic Formula and obtain

$$x = \frac{-(-2) \pm \sqrt{(-2)^2 - 4(1)(5)}}{2(1)} = \frac{2 \pm \sqrt{-16}}{2} = \frac{2 \pm 4i}{2} = 1 \pm 2i.$$

Two things are important to note. First, the zeros $1 + 2i$ and $1 - 2i$ are complex conjugates. If ever we obtain non-real zeros to a quadratic function with *real number* coefficients, the zeros will be a complex conjugate pair. (Do you see why?)

We could ask if all of the theory of polynomial division holds for non-real zeros, in particular the division algorithm and the Remainder and Factor Theorems. The answer is ‘yes.’

$$\begin{array}{r|rrr} 1+2i & 1 & -2 & 5 \\ & \downarrow & 1+2i & -5 \\ \hline & 1 & -1+2i & \boxed{0} \end{array}$$

Indeed, the above shows $x^2 - 2x + 5 = (x - [1 + 2i])(x - 1 + 2i) = (x - [1 + 2i])(x - [1 - 2i])$ which demonstrates both $(x - [1 + 2i])$ and $(x - [1 - 2i])$ are factors of $x^2 - 2x + 5$.*

But how do we know if a general polynomial has any complex zeros at all? We have many examples of polynomials with no real zeros. Can there be polynomials with no zeros whatsoever? The answer to that last question is “No.” and the theorem which provides that answer is The Fundamental Theorem of Algebra.

Theorem B.1.4 (The Fundamental Theorem of Algebra)

Suppose f is a polynomial function with complex number coefficients of degree $n \geq 1$, then f has at least one complex zero.

The Fundamental Theorem of Algebra is an example of an ‘existence’ theorem in Mathematics. Like the Intermediate Value Theorem, the Fundamental Theorem of Algebra guarantees the existence of at least one zero, but gives us no algorithm to use in finding it. In fact, as we mentioned previously, there are polynomials whose real zeros, though they exist, cannot be expressed using the ‘usual’ combinations of arithmetic symbols, and must be approximated. It took mathematicians literally hundreds of years to prove the theorem in its full generality,[†] and some of that history is recorded . Note that the Fundamental Theorem of Algebra applies to not only polynomial functions with real coefficients, but to those with complex number coefficients as well.

Suppose f is a polynomial function of degree $n \geq 1$. The Fundamental Theorem of Algebra guarantees us at least one complex zero, z_1 . The Factor Theorem guarantees that $f(x)$ factors as $f(x) = (x - z_1)q_1(x)$ for a polynomial function q_1 , which has degree $n - 1$. If $n - 1 \geq 1$, then the Fundamental Theorem of Algebra guarantees a complex zero of q_1 as well, say z_2 , so then the Factor Theorem gives us $q_1(x) = (x - z_2)q_2(x)$, and hence $f(x) = (x - z_1)(x - z_2)q_2(x)$. We can continue this process exactly n times, at which point

*It is a good review of the algebra of complex numbers to start with $(x - [1 + 2i])(x - [1 - 2i])$, perform the indicated operations, and simplify the result to $x^2 - 2x + 5$. See part 6 of Example B.2.1.

[†]So if its profound nature and beautiful subtlety escape you, no worries!

our quotient polynomial q_n has degree 0 so it's a constant. This constant is none-other than the leading coefficient of f which is carried down line by line each time we divide by factors of the form $x - c$.

Theorem B.1.5 (Complex Factorization Theorem)

Suppose f is a polynomial function with complex number coefficients. If the degree of f is n and $n \geq 1$, then f has exactly n complex zeros, counting multiplicity. If z_1, z_2, \dots, z_k are the distinct zeros of f , with multiplicities m_1, m_2, \dots, m_k , respectively, then $f(x) = a(x - z_1)^{m_1}(x - z_2)^{m_2} \cdots (x - z_k)^{m_k}$.

Theorem B.1.5 says two important things: first, every polynomial is a product of linear factors; second, every polynomial function is completely determined by its zeros, their multiplicities, and its leading coefficient. We put this theorem to good use in the next example.

Example B.1.6: Let $f(x) = 12x^5 - 20x^4 + 19x^3 - 6x^2 - 2x + 1$.

1. Find all of the complex zeros of f and state their multiplicities.
2. Factor $f(x)$ using Theorem B.1.5

Solution:

1. Since f is a fifth degree polynomial, we know that we need to perform at least three successful divisions to get the quotient down to a quadratic function. At that point, we can find the remaining zeros using the Quadratic Formula, if necessary. Using the techniques of synthetic division:

$$\begin{array}{c|ccccccc} \frac{1}{2} & 12 & -20 & 19 & -6 & -2 & 1 \\ \downarrow & 6 & -7 & 6 & 0 & -1 \\ \hline \frac{1}{2} & 12 & -14 & 12 & 0 & -2 & 0 \\ \downarrow & 6 & -4 & 4 & 2 \\ \hline -\frac{1}{3} & 12 & -8 & 8 & 4 & 0 \\ \downarrow & -4 & 4 & -4 \\ \hline 12 & -12 & 12 & 0 \end{array}$$

Our quotient is $12x^2 - 12x + 12$, whose zeros we find to be $\frac{1 \pm i\sqrt{3}}{2}$. From Theorem B.1.5, we know f has exactly 5 zeros, counting multiplicities, and as such we have the zero $\frac{1}{2}$ with multiplicity 2, and the zeros $-\frac{1}{3}, \frac{1+i\sqrt{3}}{2}$ and $\frac{1-i\sqrt{3}}{2}$, each of multiplicity 1.

2. Applying Theorem B.1.5, we are guaranteed that f factors as

$$f(x) = 12 \left(x - \frac{1}{2} \right)^2 \left(x + \frac{1}{3} \right) \left(x - \left[\frac{1+i\sqrt{3}}{2} \right] \right) \left(x - \left[\frac{1-i\sqrt{3}}{2} \right] \right)$$



A true test of Theorem B.1.5 would be to take the factored form of $f(x)$ in the previous example and multiply it out* to see that it really does reduce to $f(x) = 12x^5 - 20x^4 + 19x^3 - 6x^2 - 2x + 1$. When factoring a polynomial using Theorem B.1.5, we say that it is *factored completely over the complex numbers*, meaning that it is impossible to factor the polynomial any further using complex numbers. If we wanted to completely factor $f(x)$ over the *real numbers* then we would have stopped short of finding the nonreal zeros of f and factored f using our work from the synthetic division to write $f(x) = \left(x - \frac{1}{2}\right)^2 \left(x + \frac{1}{3}\right) (12x^2 - 12x + 12)$, or $f(x) = 12 \left(x - \frac{1}{2}\right)^2 \left(x + \frac{1}{3}\right) (x^2 - x + 1)$. Since the zeros of $x^2 - x + 1$ are nonreal, we call $x^2 - x + 1$ an *irreducible quadratic* meaning it is impossible to break it down any further using *real numbers*.

The last two results of the section show us that, theoretically, the non-real zeros of polynomial functions with real number coefficients come exclusively from irreducible quadratics.

Theorem B.1.6 (Conjugate Pairs Theorem)

If f is a polynomial function with real number coefficients and z is a complex zero of f , then so is \bar{z} .

To prove the theorem, let $f(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_2 x^2 + a_1 x + a_0$ be a polynomial function with real number coefficients. If z is a zero of f , then $f(z) = 0$, which means $a_n z^n + a_{n-1} z^{n-1} + \dots + a_2 z^2 + a_1 z + a_0 = 0$. Next, we consider $f(\bar{z})$ and apply Theorem B.2.1 below.

$$\begin{aligned}
 f(\bar{z}) &= a_n (\bar{z})^n + a_{n-1} (\bar{z})^{n-1} + \dots + a_2 (\bar{z})^2 + a_1 \bar{z} + a_0 \\
 &= a_n \bar{z}^n + a_{n-1} \bar{z}^{n-1} + \dots + a_2 \bar{z}^2 + a_1 \bar{z} + a_0 && \text{since } (\bar{z})^n = \bar{z}^n \\
 &= \overline{a_n z^n} + \overline{a_{n-1} z^{n-1}} + \dots + \overline{a_2 z^2} + \overline{a_1 z} + \overline{a_0} && \text{since the coefficients are real} \\
 &= \overline{a_n z^n} + \overline{a_{n-1} z^{n-1}} + \dots + \overline{a_2 z^2} + \overline{a_1 z} + \overline{a_0} && \text{since } \bar{z} \bar{w} = \bar{z} \bar{w} \\
 &= \overline{a_n z^n + a_{n-1} z^{n-1} + \dots + a_2 z^2 + a_1 z + a_0} && \text{since } \bar{z} + \bar{w} = \bar{z + w} \\
 &= \overline{f(z)} \\
 &= \bar{0} \\
 &= 0
 \end{aligned}$$

This shows that \bar{z} is a zero of f . So, if f is a polynomial function with real number coefficients, Theorem B.1.6 tells us that if $a + bi$ is a nonreal zero of f , then so is $a - bi$. In other words, nonreal zeros of f come in conjugate pairs. The Factor Theorem kicks in to give us both $(x - [a+bi])$ and $(x - [a-bi])$ as factors of $f(x)$ which means $(x - [a+bi])(x - [a-bi]) = x^2 + 2ax + (a^2 + b^2)$ is an irreducible quadratic factor of f . As a result, we have our last theorem of the section.

*This is a good chance to test your algebraic mettle and see that all of this does actually work.

Theorem B.1.7 (Real Factorization Theorem)

Suppose f is a polynomial function with real number coefficients. Then $f(x)$ can be factored into a product of linear factors corresponding to the real zeros of f and irreducible quadratic factors which give the nonreal zeros of f .

Example B.1.7: Let $f(x) = x^4 + 64$.

1. Use synthetic division to show that $x = 2 + 2i$ is a zero of f .
2. Find the remaining complex zeros of f .
3. Completely factor $f(x)$ over the complex numbers.
4. Completely factor $f(x)$ over the real numbers.

Solution:

1. Remembering to insert the 0's in the synthetic division tableau we have

$$\begin{array}{c|ccccc} 2+2i & 1 & 0 & 0 & 0 & 64 \\ \downarrow & 2+2i & 8i & -16+16i & -64 \\ \hline 1 & 2+2i & 8i & -16+16i & \boxed{0} \end{array}$$

2. Since f is a fourth degree polynomial, we need to make two successful divisions to get a quadratic quotient. Since $2 + 2i$ is a zero, we know from Theorem B.1.6 that $2 - 2i$ is also a zero. We continue our synthetic division tableau.

$$\begin{array}{c|ccccc} 2+2i & 1 & 0 & 0 & 0 & 64 \\ \downarrow & 2+2i & 8i & -16+16i & -64 \\ \hline 2-2i & 1 & 2+2i & 8i & -16+16i & \boxed{0} \\ \downarrow & 2-2i & 8-8i & 16-16i & \\ \hline 1 & 4 & 8 & \boxed{0} \end{array}$$

Our quotient polynomial is $x^2 + 4x + 8$. Using the quadratic formula, we solve $x^2 + 4x + 8 = 0$ and find the remaining zeros are $-2 + 2i$ and $-2 - 2i$.

3. Using Theorem B.1.5, we get $f(x) = (x - [2 - 2i])(x - [2 + 2i])(x - [-2 + 2i])(x - [-2 - 2i])$.
4. To find the irreducible quadratic factors of $f(x)$, we multiply the factors together which correspond to the conjugate pairs. We find $(x - [2 - 2i])(x - [2 + 2i]) = x^2 - 4x + 8$, and $(x - [-2 + 2i])(x - [-2 - 2i]) = x^2 + 4x + 8$, so $f(x) = (x^2 - 4x + 8)(x^2 + 4x + 8)$.

We close this section with an example where we are asked to manufacture a polynomial function with certain characteristics.

Example B.1.8: Find a polynomial function p of lowest degree that has integer coefficients and satisfies all of the following criteria:

- the graph of $y = p(x)$ touches and rebounds from the x -axis at $(\frac{1}{3}, 0)$
- $x = 3i$ is a zero of p .
- as $x \rightarrow -\infty$, $p(x) \rightarrow -\infty$
- as $x \rightarrow \infty$, $p(x) \rightarrow -\infty$

Solution:

To solve this problem, we will need a good understanding of the relationship between the x -intercepts of the graph of a function and the zeros of a function, the Factor Theorem, the role of multiplicity, complex conjugates, the Complex Factorization Theorem, and end behavior of polynomial functions. (In short, you'll need most of the major concepts of this chapter.) Since the graph of p touches the x -axis at $(\frac{1}{3}, 0)$, we know $x = \frac{1}{3}$ is a zero of even multiplicity. Since we are after a polynomial of lowest degree, we need $x = \frac{1}{3}$ to have multiplicity exactly 2. The Factor Theorem now tells us $(x - \frac{1}{3})^2$ is a factor of $p(x)$. Since $x = 3i$ is a zero and our final answer is to have integer (hence, real) coefficients, $x = -3i$ is also a zero. The Factor Theorem kicks in again to give us $(x - 3i)$ and $(x + 3i)$ as factors of $p(x)$. We are given no further information about zeros or intercepts so we conclude, by the Complex Factorization Theorem that $p(x) = a(x - \frac{1}{3})^2(x - 3i)(x + 3i)$ for some real number a . Expanding this, we get $p(x) = ax^4 - \frac{2a}{3}x^3 + \frac{82a}{9}x^2 - 6ax + a$. In order to obtain integer coefficients, we know a must be an integer multiple of 9. Our last concern is end behavior. Since the leading term of $p(x)$ is ax^4 , we need $a < 0$ to get $p(x) \rightarrow -\infty$ as $x \rightarrow \pm\infty$. Hence, if we choose $x = -9$, we get $p(x) = -9x^4 + 6x^3 - 82x^2 + 54x - 9$. We can verify our handiwork using the techniques developed in this chapter. \square

B.2 Complex Numbers

Note: Attribution: [SZ], §A.11

The equation $x^2 + 1 = 0$ has no real number solutions. However, it *would* have solutions if we could make sense of $\sqrt{-1}$. The *Complex Numbers* do just that - they give us a mechanism for working with $\sqrt{-1}$. As such, the set of complex numbers fill in an algebraic gap left by the set of real numbers.

Here's the basic plan. There is no real number x with $x^2 = -1$, since for any real number $x^2 \geq 0$. However, we could formally extract square roots and write $x = \pm\sqrt{-1}$. We build the complex numbers by relabeling the quantity $\sqrt{-1}$ as i , the unfortunately misnamed *imaginary unit*.^{*} The number i , while not a real number, is defined so that it plays along well with real numbers and acts very much like any other radical expression. For instance, $3(2i) = 6i$, $7i - 3i = 4i$, $(2 - 7i) + (3 + 4i) = 5 - 3i$, and so forth. The key properties which distinguish i from the real numbers are listed below.

Definition B.2.1

The imaginary unit i satisfies the two following properties:

1. $i^2 = -1$
2. If c is a real number with $c \geq 0$ then $\sqrt{-c} = i\sqrt{c}$

Property 1 in the previous definition establishes that i does act as a square root[†] of -1 , and property 2 establishes what we mean by the ‘principal square root’ of a negative real number. In property 2, it is important to remember the restriction on c . For example, it is perfectly acceptable to say $\sqrt{-4} = i\sqrt{4} = i(2) = 2i$. However, $\sqrt{-(-4)} \neq i\sqrt{-4}$, otherwise, we'd get

$$2 = \sqrt{4} = \sqrt{-(-4)} = i\sqrt{-4} = i(2i) = 2i^2 = 2(-1) = -2,$$

which is unacceptable. The moral of this story is that the general properties of radicals do not apply for even roots of negative quantities. With Definition B.2.1 in place, we can define the set of *complex numbers*.

A *complex number* is a number of the form $a + bi$, where a and b are real numbers and i is the imaginary unit. The set of complex numbers is denoted \mathbb{C} .

Complex numbers include things you'd normally expect, like $3 + 2i$ and $\frac{2}{5} - i\sqrt{3}$. However, don't forget that a or b could be zero, which means numbers like $3i$ and 6 are also complex numbers. In other words, don't forget that the complex numbers *include* the real numbers,[‡] so 0 and $\pi - \sqrt{2}i$ are both considered complex numbers. The arithmetic of complex numbers

*Some Technical Mathematics textbooks label it ‘ j ’. While it carries the adjective ‘imaginary’, these numbers have essential real-world implications. For example, every electronic device owes its existence to the study of ‘imaginary’ numbers.

[†]Note the use of the indefinite article ‘ a ’. Whatever beast is chosen to be i , $-i$ is the other square root of -1 .

[‡]In the language of set notation, $\mathbb{R} \subseteq \mathbb{C}$.

is as you would expect. The only things you need to remember are the two properties above. The next example should help recall how these animals behave.

Example B.2.1: Perform the indicated operations.

$$\begin{array}{lll} 1. (1 - 2i) - (3 + 4i) & 2. (1 - 2i)(3 + 4i) & 3. \frac{1 - 2i}{3 - 4i} \\ 4. \sqrt{-3}\sqrt{-12} & 5. \sqrt{(-3)(-12)} & 6. (x - [1 + 2i])(x - [1 - 2i]) \end{array}$$

Solution:

1. As mentioned earlier, we treat expressions involving i as we would any other radical. We distribute and combine like terms:

$$\begin{aligned} (1 - 2i) - (3 + 4i) &= 1 - 2i - 3 - 4i && \text{Distribute} \\ &= -2 - 6i && \text{Gather like terms} \end{aligned}$$

Technically, we'd have to rewrite our answer $-2 - 6i$ as $(-2) + (-6)i$ to be (in the strictest sense) ‘in the form $a + bi$ ’. That being said, even pedants have their limits, so $-2 - 6i$ is good enough.

2. Using the Distributive Property (a.k.a. F.O.I.L.), we get

$$\begin{aligned} (1 - 2i)(3 + 4i) &= (1)(3) + (1)(4i) - (2i)(3) - (2i)(4i) && \text{F.O.I.L.} \\ &= 3 + 4i - 6i - 8i^2 \\ &= 3 - 2i - 8(-1) && i^2 = -1 \\ &= 3 - 2i + 8 \\ &= 11 - 2i \end{aligned}$$

3. How in the world are we supposed to simplify $\frac{1-2i}{3-4i}$? Well, we deal with the denominator $3 - 4i$ as we would any other denominator containing two terms, one of which is a square root. We multiply both numerator and denominator by $3 + 4i$, the (complex) conjugate of $3 - 4i$. Doing so produces

$$\begin{aligned} \frac{1 - 2i}{3 - 4i} &= \frac{(1 - 2i)(3 + 4i)}{(3 - 4i)(3 + 4i)} && \text{Equivalent Fractions} \\ &= \frac{3 + 4i - 6i - 8i^2}{9 - 16i^2} && \text{F.O.I.L.} \\ &= \frac{3 - 2i - 8(-1)}{9 - 16(-1)} && i^2 = -1 \\ &= \frac{11 - 2i}{25} \\ &= \frac{11}{25} - \frac{2}{25}i \end{aligned}$$

4. We use property 2 of Definition B.2.1 first, then apply the rules of radicals applicable to real numbers to get $\sqrt{-3}\sqrt{-12} = (i\sqrt{3})(i\sqrt{12}) = i^2\sqrt{3 \cdot 12} = -\sqrt{36} = -6$.

5. We adhere to the order of operations here and perform the multiplication before the radical to get $\sqrt{(-3)(-12)} = \sqrt{36} = 6$.
6. We brute force multiply using the distributive property and find that

$$\begin{aligned}
 (x - [1 + 2i])(x - [1 - 2i]) &= x^2 - x[1 - 2i] - x[1 + 2i] + [1 - 2i][1 + 2i] \\
 &= x^2 - x + 2ix - x - 2ix + 1 - 2i + 2i - 4i^2 \\
 &= x^2 - 2x + 1 - 4(-1) \\
 &= x^2 - 2x + 5
 \end{aligned}$$

□

In the previous example, we used the ‘conjugate’ idea from simplifying radical equations to divide two complex numbers. More generally, the *complex conjugate* of a complex number $a + bi$ is the number $a - bi$. The notation commonly used for complex conjugation is a ‘bar’: $\overline{a + bi} = a - bi$. For example, $\overline{3 + 2i} = 3 - 2i$ and $\overline{3 - 2i} = 3 + 2i$. To find $\overline{6}$, we note that $\overline{6} = \overline{6 + 0i} = 6 - 0i = 6$, so $\overline{6} = 6$. Similarly, $\overline{4i} = -4i$, since $\overline{4i} = \overline{0 + 4i} = 0 - 4i = -4i$. Note that $\overline{3 + \sqrt{5}} = 3 + \sqrt{5}$, not $3 - \sqrt{5}$, since $\overline{3 + \sqrt{5}} = \overline{3 + \sqrt{5} + 0i} = 3 + \sqrt{5} - 0i = 3 + \sqrt{5}$. Here, the conjugation specified by the ‘bar’ notation involves reversing the sign before $i = \sqrt{-1}$, not before $\sqrt{5}$. The properties of the conjugate are summarized in the following theorem.

Theorem B.2.1 (Properties of the Complex Conjugate)

Let z and w be complex numbers.

- $\overline{\overline{z}} = z$
- $\overline{z + w} = \overline{z} + \overline{w}$
- $\overline{zw} = \overline{z}\overline{w}$
- $\overline{z^n} = (\overline{z})^n$, for any natural number n
- z is a real number if and only if $\overline{z} = z$.

Theorem B.2.1 says in part that complex conjugation works well with addition, multiplication and powers. The proofs of these properties can best be achieved by writing out $z = a + bi$ and $w = c + di$ for real numbers a, b, c and d . Next, we compute the left and right sides of each equation and verify that they are the same.

The proof of the first property is a very quick exercise.* To prove the second property, we compare $\overline{z + w}$ with $\overline{z} + \overline{w}$. We have $\overline{z + w} = \overline{a + bi + c + di} = a - bi + c - di$. To find $\overline{z + w}$, we first compute

$$z + w = (a + bi) + (c + di) = (a + c) + (b + d)i$$

so

$$\overline{z + w} = \overline{(a + c) + (b + d)i} = (a + c) - (b + d)i = a + c - bi - di = a - bi + c - di = \overline{z} + \overline{w}$$

*Trust us on this.

As such, we have established $\overline{z+w} = \overline{z} + \overline{w}$. The proof for multiplication works similarly. The proof that the conjugate works well with powers can be viewed as a repeated application of the product rule, and is best proved using a technique called Mathematical Induction. The last property is a characterization of real numbers. If z is real, then $z = a+0i$, so $\overline{z} = a-0i = a = z$. On the other hand, if $z = \overline{z}$, then $a+bi = a-bi$ which means $b = -b$ so $b = 0$. Hence, $z = a+0i = a$ and is real.

We now return to the business of solving quadratic equations. Consider $x^2 - 2x + 5 = 0$. The discriminant $b^2 - 4ac = -16$ is negative, so we know that there are no *real* solutions, since the Quadratic Formula would involve the term $\sqrt{-16}$. Complex numbers, however, are built just for such situations, so we can go ahead and apply the Quadratic Formula to get:

$$x = \frac{-(-2) \pm \sqrt{(-2)^2 - 4(1)(5)}}{2(1)} = \frac{2 \pm \sqrt{-16}}{2} = \frac{2 \pm 4i}{2} = 1 \pm 2i.$$

Example B.2.2: Find the complex solutions to the following equations.*

$$\begin{array}{lll} 1. \frac{2x}{x+1} = x+3 & 2. 2t^4 = 9t^2 + 5 & 3. z^3 + 1 = 0 \end{array}$$

Solution:

- Clearing fractions yields a quadratic equation so we then proceed via normal quadratic equation methods.

$$\begin{aligned} \frac{2x}{x+1} &= x+3 \\ 2x &= (x+3)(x+1) && \text{Multiply by } (x+1) \text{ to clear denominators} \\ 2x &= x^2 + x + 3x + 3 && \text{F.O.I.L.} \\ 2x &= x^2 + 4x + 3 && \text{Gather like terms} \\ 0 &= x^2 + 2x + 3 && \text{Subtract } 2x \end{aligned}$$

From here, we apply the Quadratic Formula

$$\begin{aligned} x &= \frac{-2 \pm \sqrt{2^2 - 4(1)(3)}}{2(1)} && \text{Quadratic Formula} \\ &= \frac{-2 \pm \sqrt{-8}}{2} && \text{Simplify} \\ &= \frac{-2 \pm i\sqrt{8}}{2} && \text{Definition of } i \\ &= \frac{-2 \pm i2\sqrt{2}}{2} && \text{Product Rule for Radicals} \\ &= \frac{2(-1 \pm i\sqrt{2})}{2} && \text{Factor and reduce} \\ &= -1 \pm i\sqrt{2} \end{aligned}$$

We get two answers: $x = -1 + i\sqrt{2}$ and its conjugate $x = -1 - i\sqrt{2}$. Checking both of these answers reviews all of the salient points about complex number arithmetic and is therefore strongly encouraged.

*Remember, all real numbers are complex numbers, so ‘complex solutions’ means both real and non-real answers.

2. Since we have three terms, and the exponent on one term ('4' on t^4) is exactly twice the exponent on the other ('2' on t^2), we have a Quadratic in Disguise. We proceed accordingly.

$$\begin{array}{rcl} 2t^4 & = & 9t^2 + 5 \\ 2t^4 - 9t^2 - 5 & = & 0 \quad \text{Subtract } 9t^2 \text{ and } 5 \\ (2t^2 + 1)(t^2 - 5) & = & 0 \quad \text{Factor} \\ 2t^2 + 1 = 0 \quad \text{or} \quad t^2 = 5 & & \text{Zero Product Property} \end{array}$$

From $2t^2 + 1 = 0$ we get $2t^2 = -1$, or $t^2 = -\frac{1}{2}$. We extract square roots as follows:

$$t = \pm \sqrt{-\frac{1}{2}} = \pm i \sqrt{\frac{1}{2}} = \pm i \frac{\sqrt{1}}{\sqrt{2}} = \pm i \frac{1}{\sqrt{2}} = \pm \frac{i\sqrt{2}}{2},$$

where we have rationalized the denominator per convention. From $t^2 = 5$, we get $t = \pm\sqrt{5}$. In total, we have four complex solutions - two real: $t = \pm\sqrt{5}$ and two non-real: $t = \pm\frac{i\sqrt{2}}{2}$.

3. To find the *real* solutions to $z^3 + 1 = 0$, we can subtract the 1 from both sides and extract cube roots: $z^3 = -1$, so $z = \sqrt[3]{-1} = -1$. It turns out there are two more non-real complex number solutions to this equation. To get at these, we factor:

$$\begin{array}{rcl} z^3 + 1 & = & 0 \\ (z + 1)(z^2 - z + 1) & = & 0 \quad \text{Factor (Sum of Two Cubes)} \\ z + 1 = 0 \quad \text{or} \quad z^2 - z + 1 = 0 & & \end{array}$$

From $z + 1 = 0$, we get our real solution $z = -1$. From $z^2 - z + 1 = 0$, we apply the Quadratic Formula to get:

$$z = \frac{-(-1) \pm \sqrt{(-1)^2 - 4(1)(1)}}{2(1)} = \frac{1 \pm \sqrt{-3}}{2} = \frac{1 \pm i\sqrt{3}}{2}$$

Thus we get *three* solutions to $z^3 + 1 = 0$ - one real: $z = -1$ and two non-real: $z = \frac{1 \pm i\sqrt{3}}{2}$. As always, the reader is encouraged to test their algebraic mettle and check these solutions.

□

It is no coincidence that the non-real solutions to the equations in Example B.2.2 appear in complex conjugate pairs. Any time we use the Quadratic Formula to solve an equation with real coefficients, the answers will form a complex conjugate pair owing to the \pm in the Quadratic Formula.

Theorem B.2.2 (Discriminant Theorem)

Given a Quadratic Equation $ax^2 + bx + c = 0$, where a , b and c are real numbers, let $D = b^2 - 4ac$ be the discriminant.

- If $D > 0$, there are two distinct real number solutions to the equation.
- If $D = 0$, there is one (repeated) real number solution.
‘Repeated’ here comes from the fact that ‘both’ solutions $\frac{-b \pm 0}{2a}$ reduce to $-\frac{b}{2a}$.
- If $D < 0$, there are two non-real solutions which form a complex conjugate pair.

B.3 Differentiation and Integration Techniques

In this section, we will cover some of the basic derivative and integral formulas that will be necessary for success in Differential Equations. In order to be able to deal with equations that involve derivatives, we need to be able to take derivatives as well as remove them.

B.3.1 Derivative and Integral Formulas

The following is a table of some of the basic derivative formulas covered in a Calculus 1 course.

Function $f(x)$	Derivative $f'(x)$
x^n any n	nx^{n-1}
$\ln(x)$	$\frac{1}{x} = x^{-1}$
C constant	0
e^x	e^x
e^{ax}	ae^{ax}
$\sin(x)$	$\cos(x)$
$\cos(x)$	$-\sin(x)$
$\tan(x)$	$\sec^2(x)$
$\arctan(x) = \tan^{-1}(x)$	$\frac{1}{x^2+1}$

Similarly, we have a table for some basic integral formulas. As integration is the inverse operation to differentiation, this table will look like the reverse version of the previous table.

Function $f(x)$	Integral $\int f(x) dx$
x^n any $n \neq -1$	$\frac{1}{n+1}x^{n+1} + C$
$\frac{1}{x}$	$\ln(x) + C$
e^x	$e^x + C$
e^{ax}	$\frac{1}{a}e^{ax} + C$
$\sin(x)$	$-\cos(x) + C$
$\cos(x)$	$\sin(x) + C$
$\frac{1}{x^2+1}$	$\arctan(x) + C$ or $\tan^{-1}(x) + C$

B.3.2 Derivative Rules

The tables above only list a few simple functions for which we know how to compute the derivative and integral. However, there are some nice properties of derivatives and integrals that make this enough for our needs.

Linearity of the Derivative and Integral

The derivative and integral are both linear operators. This means that if we have two functions $f(x)$ and $g(x)$, and two constants a and b , then

$$\frac{d}{dx} (af(x) + bg(x)) = a\frac{df}{dx} + b\frac{dg}{dx}.$$

That is, we can move constants and addition and subtractions out of the differentiation, reducing a complicated function down to simpler functions that we know how to differentiate.

The same is true for integration or antiderivatives; if we have functions $f(x)$ and $g(x)$ and constants a and b , then

$$\int af(x) + bg(x) \, dx = a \int f(x) \, dx + b \int g(x) \, dx.$$

Example B.3.1: Compute the following derivatives and integrals using linearity and the table of known formulas.

1. $\frac{d}{dx} (x^3 + \frac{4}{x^2} + 3e^x)$
2. $\frac{d}{dx} (\sin(x) - 2\cos(x) + 5\ln(x))$
3. $\int \frac{2x^3+4x}{x^2} \, dx$
4. $\int 2\cos(x) - \frac{3}{x^2+1} \, dx$

Solution:

1. For this, we can use linearity and our formulas to write

$$\begin{aligned} \frac{d}{dx} \left(x^3 + \frac{4}{x^2} + 3e^x \right) &= \frac{d}{dx} (x^3) + \frac{d}{dx} \left(\frac{4}{x^2} \right) + \frac{d}{dx} (3e^x) \\ &= \frac{d}{dx} (x^3) + 4 \frac{d}{dx} (x^{-2}) + 3 \frac{d}{dx} (e^x) \\ &= 3x^2 - 8x^{-3} + 3e^x. \end{aligned}$$

2. This one gives

$$\begin{aligned} \frac{d}{dx} (\sin(x) - 2\cos(x) + 5\ln(x)) &= \frac{d}{dx} (\sin(x)) - \frac{d}{dx} (2\cos(x)) + \frac{d}{dx} (5\ln(x)) \\ &= \frac{d}{dx} (\sin(x)) - 2 \frac{d}{dx} (\cos(x)) + 5 \frac{d}{dx} (\ln(x)) \\ &= \cos(x) + 2\sin(x) + \frac{5}{x}. \end{aligned}$$

3. For this problem, we first want to simplify the expression algebraically, then integrate each term using linearity.

$$\begin{aligned} \int \frac{2x^3+4x}{x^2} \, dx &= \int \frac{2x^3}{x^2} + \frac{4x}{x^2} \, dx \\ &= \int 2x + \frac{4}{x} \, dx \\ &= 2 \int x \, dx + 4 \int \frac{1}{x} \, dx \\ &= x^2 + 4\ln(|x|) + C. \end{aligned}$$

4. This problem uses standard linearity to get to the final answer.

$$\begin{aligned}\int 2 \cos(x) - \frac{3}{x^2 + 1} dx &= 2 \int \cos(x) dx - 3 \int \frac{1}{x^2 + 1} dx \\ &= 2 \sin(x) - 3 \arctan(x) + C.\end{aligned}$$

]

Product and Quotient Rule

Linearity gives us a way to handle sums and differences of derivatives. What about products? It turns out that doesn't work as simply, but there is still a nice formula to work it out. This gives us the Product Rule. If we have two functions $f(x)$ and $g(x)$, then

$$\frac{d}{dx} (f(x)g(x)) = f(x)\frac{dg}{dx} + \frac{df}{dx}g(x).$$

That is, the derivative has two terms, the first function times the derivative of the second, and the derivative of the first function times the second function. The product rule can also be used to compute the product of more than two functions; the general formula is that only one function is differentiated at a time and each function should be differentiated once. That is, for three functions, the formula is

$$\frac{d}{dx} (f(x)g(x)h(x)) = \frac{df}{dx}g(x)h(x) + f(x)\frac{dg}{dx}h(x) + f(x)g(x)\frac{dh}{dx}.$$

The Quotient Rule gives us a way to do the same thing, but with quotients. The formula here is that

$$\frac{d}{dx} \left(\frac{f(x)}{g(x)} \right) = \frac{g(x)\frac{df}{dx} - f(x)\frac{dg}{dx}}{(g(x))^2}.$$

This can also be derived using the product rule and the chain rule. It is important to get the order of the numerator correct, as there is a subtraction on top. For the product rule, the addition means that the order doesn't matter, but if the order for the quotient rule is incorrect, there will be an additional minus sign in the answer.

Example B.3.2: Compute the following derivatives.

$$1. \frac{d}{dx} (e^x \cos(x))$$

$$2. \frac{d}{dx} \left(\frac{\sin(x)}{x^2} \right)$$

$$3. \frac{d}{dx} \left(\frac{x^3 e^x}{\tan(x)} \right).$$

Solution:

1. This is a direct application of the product rule.

$$\begin{aligned}\frac{d}{dx} (e^x \cos(x)) &= e^x \frac{d}{dx} (\cos(x)) + \frac{d}{dx} (e^x) \cos(x) \\ &= e^x(-\sin(x)) + (e^x) \cos(x) \\ &= e^x(\cos(x) - \sin(x)).\end{aligned}$$

2. This is a direct application of the quotient rule.

$$\begin{aligned}\frac{d}{dx} \left(\frac{\sin(x)}{x^2} \right) &= \frac{x^2 \frac{d}{dx}(\sin(x)) - \sin(x) \frac{d}{dx}(x^2)}{(x^2)^2} \\ &= \frac{x^2 \cos(x) - \sin(x)(2x)}{x^4} \\ &= \frac{x \cos(x) - 2 \sin(x)}{x^3}.\end{aligned}$$

3. For this problem, we need to apply both the product rule and the quotient rule. Since the quotient rule is on the outside, we apply it first.

$$\begin{aligned}\frac{d}{dx} \left(\frac{x^3 e^x}{\tan(x)} \right) &= \frac{\tan(x) \frac{d}{dx}(x^3 e^x) - x^3 e^x \frac{d}{dx}(\tan(x))}{(\tan(x))^2} \\ &= \frac{\tan(x) \left(x^3 \frac{d}{dx}(e^x) + \frac{d}{dx}(x^3) e^x \right) - x^3 e^x \sec^2(x)}{\tan^2(x)} \\ &= \frac{\tan(x) (x^3 e^x + 3x^2 e^x) - x^3 e^x \sec^2(x)}{\tan^2(x)} \\ &= \frac{e^x (x^3 + 3x^2)}{\tan(x)} - \frac{x^3 e^x}{\sin^2(x)}.\end{aligned}$$

]

Chain Rule

The only type of function we haven't discussed yet for differentiation is composite functions, and that is handled by the Chain Rule. For example, we don't have a direct way (yet) to differentiate functions like $\sin(3x)$ or $\frac{1}{x^3+4x+1}$, and the Chain Rule lets us do that. This rule tells us that, for functions $f(x)$ and $g(x)$, we can compute the derivative of the composition $(f \circ g)(x)$ or $f(g(x))$ is

$$\frac{d}{dx}(f(g(x))) = f'(g(x))g'(x).$$

This means that we differentiate the “outside” function f , plug in the inside function, and then multiply this by the derivative of the “inside” function g . It requires us to identify what the “inner” and “outer” functions are, and then the formula gives what the derivative should be. This can be done in a few different ways, either moving from outside in, or moving from inside out. These problems are conventionally written with $u(x)$ as the inside function, but any letter can be used.

Example B.3.3: Compute the derivative of each of the following functions.

1. $f_1(x) = (x^3 + 5x + 1)^5$
2. $f_2(x) = \cos(3x^2 + 1)$
3. $f_3(x) = (1 + \sin(3x))^4$

Solution:

1. For this problem, we take $f(u) = u^5$ and $u(x) = x^3 + 5x + 1$, which gives that composing these functions gives the f_1 that we started with. Therefore, since $f'(u) = 5u^4$ and $u'(x) = 3x^2 + 5$, we have that

$$f'_1(x) = f'(u)u'(x) = 5u^4(3x^2 + 5) = 5(x^3 + 5x + 1)^4(3x^2 + 5).$$

2. For this case, the outside function is $\cos(u)$ and the inner function is $u(x) = 3x^2 + 1$. Using the same process, we get that

$$f'_2(x) = -\sin(u)(6x) = -6x \sin(3x^2 + 1).$$

3. Starting from the outside, we see that we can take $f(u) = u^4$. This makes $u(x) = 1 + \sin(3x)$, but we can't differentiate this directly; it requires another iteration of the Chain Rule. Taking $u(x) = 1 + \sin(v)$ for $v(x) = 3x$, we can then compute the derivative of each of these functions, and our original function $f_3(x) = f(u(v(x)))$. We can extend the Chain Rule to apply to three functions by taking it one step at a time. The result of this process is that

$$\frac{d}{dx}(f(u(v(x)))) = f'(u(v(x)) \frac{d}{dx}(u(v(x))) = f'(u(v(x)))u'(v(x))v'(x),$$

so you need to pull off one derivative at time to get to the correct computation. Thus, for this problem, we get that

$$f'_3(x) = 4u^3(\cos(v))(3) = 12u^3 \cos(v) = 12(1 + \sin(3x))^3 \cos(3x).$$

□

B.3.3 Integration Techniques

Another main topic that will be needed throughout study of differential equations is various integration techniques. When trying to solve questions that involve derivatives, integration will be a very important step in that process.

Substitution

The substitution method for integration serves as the inverse operation to the Chain Rule for differentiation. Since

$$\frac{d}{dx}(f(u(x))) = f'(u(x))u'(x),$$

the definition of the integral as an antiderivative gives that

$$\int f'(u(x))u'(x) dx = f(u(x)) + C.$$

Integrals of this form can be computed using this formula, but it is often easier to think of this process in terms of “changing variables.” This means the following: If we have an integral that looks like

$$\int f'(u(x))u'(x) \, dx$$

then we can define the variable u to represent the entire function $u(x)$. Then the differential du is defined by

$$du = u'(x)dx.$$

Then we can substitute both u and du into the original expression to get that

$$\int f'(u(x))u'(x) \, dx = \int f'(u) \, du = f(u) + C = f(u(x)) + C.$$

The last component of this process is changing the limits of integration if a definite integral is being computed. The idea is that an integral in x (denoted by dx) has its limits also in terms of x , whereas the du integral has endpoints given in terms of u . The main way this comes up in problems is that

$$\int_a^b f(u(x))u'(x) \, dx = \int_{u(a)}^{u(b)} f(u) \, du$$

because we know that u is written in terms of x as $u = u(x)$. Thus if we plug the x endpoints into this function, we will be the new u endpoints.

Example B.3.4: Compute the following integrals using substitution.

1. $\int \cos(4x) \, dx$
2. $\int x \sin(3x^2 + 1) \, dx$
3. $\int_0^2 \frac{3x^2}{x^3 + 4} \, dx$

Solution:

1. For this situation we want to set $u = 4x$, because then the integrand, once we make the change of variables, will be $\cos(u)$, which we know how to integrate. With this, we have $du = 4 \, dx$, which we can rewrite as $dx = \frac{1}{4} \, du$. Plugging all of this in gives that

$$\int \cos(4x) \, dx = \int \cos(u) \frac{1}{4} \, du = \frac{1}{4} \int \cos(u) \, du = \frac{1}{4} \sin(u) + C = \frac{1}{4} \sin(4x) + C.$$

2. For the same reason, we want to set $u = 3x^2 + 1$ to make the resulting integral $\sin(u) \, du$. In this case, we have $du = 6x \, dx$ or $x \, dx = \frac{1}{6} \, du$. Plugging all of this in, we get

$$\int x \sin(3x^2 + 1) \, dx = \int \sin(u) \frac{1}{6} \, du = \frac{1}{6} \int \sin(u) \, du = -\frac{1}{6} \cos(u) + C = -\frac{1}{6} \cos(3x^2 + 1) + C.$$

3. We can follow the same logic here as for the previous examples, but since we have a definite integral, we also need to switch the limits of integration. In this case, we want to pick $u = x^3 + 4$, which gives $du = 3x^2 dx$. This gives the resulting integral as

$$\int \frac{3x^2}{x^3 + 4} dx = \int \frac{1}{u} du.$$

For the limits of integration, we take the function $u(x) = x^3 + 4$ and plug in the original values of 0 and 2. This gives the value 4 and $x = 0$ and the value 12 at $x = 2$. Therefore, the result of this computation is

$$\int_0^2 \frac{3x^2}{x^3 + 4} dx = \int_4^{12} \frac{1}{u} du = \ln(|u|) \Big|_4^{12} = \ln(12) - \ln(4) = \ln(3).$$

□

There can be some cases where these techniques will not work, because the u' term that you are looking for doesn't quite appear in the expression you are trying to integrate. In cases like this, you may need to use some more complicated methods (like trigonometric substitution) or connect to inverse trigonometric integrals or other known formulas.

Integration by Parts

Integration by parts is the method used to handle integrals of a product of functions. Like the substitution method is the inverse of the Chain Rule, integration by parts is the inverse of the product rule. There are two main formulas that are used for this process. For two differentiable functions $f(x)$ and $g(x)$, we have

$$\int f(x)g'(x) dx = f(x)g(x) - \int g(x)f'(x)dx.$$

The other form is

$$\int u dv = uv - \int v du,$$

which matches the original form after setting $u = f(x)$ and $v = g(x)$.

The most important part of this process is picking the appropriate functions for u and v in this formula. The general rule is given by the following list

- Logarithmic functions
- Inverse Functions
- Algebraic or Polynomial Functions
- Trigonometric Functions (sine and cosine)
- Exponential Functions

and you want to make u , the function that you are differentiating, the one that is higher on the list. The main reason for this list is that integration is much harder than differentiation, and so we generally want to integrate the part of the product that we have a formula for. This is why logarithms and inverses are on the top; we know how to differentiate them, but integration is difficult or impossible. Polynomials are good for both differentiation and integrals, but the benefit of differentiating them is that they eventually disappear, leaving us with an integral that we know how to solve. For example, x^2 becomes $2x$, and then differentiating a second time gives 2, which is just a constant and can be removed from the integral. Trigonometric and Exponential functions are interchangeable, they are easy to differentiate and integrate, and they don't go away if we keep applying either operation.

This method can also be performed multiple times by redefining u and v and applying the same process to the integral that remains on the right-hand side. When doing this, it is important not to reverse the roles of u and v , because then the process will just undo what was done in the first step. There are also some cases where circular reasoning is used, integrating by parts twice to get to the same expression on both sides of the equal sign, which can then be solved for. One of those will be shown in the examples below.

Example B.3.5: Compute the following integrals.

1. $\int x \sin(2x) dx$
2. $\int 3x^2 e^{4x} dx$
3. $\int e^{2x} \cos(3x) dx$

Solution:

1. Based on our list, we should choose $u = x$, as it is a polynomial function. This means that $dv = \sin(2x) dx$. From this, we get that $du = dx$ and we compute v by integrating $\sin(2x) dx$, which requires a substitution. This results in $v = -\frac{1}{2} \cos(2x)$. Thus, the integration by parts formula gives

$$\int x \sin(2x) dx = x \left(-\frac{1}{2} \cos(2x) \right) - \int \left(-\frac{1}{2} \cos(2x) \right) dx.$$

This last integral we can compute directly, again requiring a substitution. Thus, the final answer is

$$\int x \sin(2x) dx = -\frac{x}{2} \cos(2x) + \frac{1}{4} \sin(2x) + C.$$

2. By the same argument as the first example, we want to pick $u = 3x^2$ so then $dv = e^{4x} dx$. We can then compute that $du = 6x dx$ and $v = \frac{1}{4} e^{4x}$. Thus, integration by parts gives

$$\int 3x^2 e^{4x} dx = 3x^2 \left(\frac{1}{4} e^{4x} \right) - \int \left(\frac{1}{4} e^{4x} \right) (6x dx) = \frac{3}{4} x^2 e^{4x} - \int \frac{3}{2} x e^{4x} dx.$$

This last integral is not something that we know how to compute. However, it looks like a product, so we should be able to work it out using integration by parts. We can

set $u = \frac{3}{2}x$ and $dv = e^{4x} dx$. This is the same dv as before, which is good. If we had picked $dv = \frac{3}{2}x dx$, we would have just gotten back to where we started. From these choices, we get that $du = \frac{3}{2} dx$ and $v = \frac{1}{4}e^{4x}$. Integration by parts then gives that

$$\int \frac{3}{2}xe^{4x} dx = \frac{3}{8}xe^{4x} - \int \frac{3}{8}e^{4x} dx.$$

Now we can compute this last integral, which will give another factor of $\frac{1}{4}$, resulting in

$$\int \frac{3}{2}xe^{4x} dx = \frac{3}{8}xe^{4x} - \frac{3}{32}e^{4x} + C.$$

Finally, we can combine this with our first integration by parts step to get that

$$\int 3x^2e^{4x} dx = \frac{3}{4}x^2e^{4x} - \frac{3}{8}xe^{4x} + \frac{3}{32}e^{4x} + C.$$

3. For this example, we have both an exponential and a trigonometric function. We can pick either one to be u and dv , and as long as we are consistent with that choice, we will get to the correct answer. For this, we will choose $u = e^{2x}$ and $dv = \cos(3x) dx$. From these, we can compute that $du = 2e^{2x} dx$ and $v = \frac{1}{3}\sin(3x)$. Thus, integration by parts tells us that

$$\int e^{2x} \cos(3x) dx = \frac{1}{3}e^{2x} \sin(3x) - \int \frac{2}{3}e^{2x} \sin(3x) dx.$$

This new integral is again a product, so we need to handle it using integration by parts. To do this, we are going to pick $u = \frac{2}{3}e^{2x}$ and $dv = \sin(3x) dx$. *Note:* If you pick $u = \sin(3x)$ and $dv = \frac{2}{3}e^{2x} dx$, the second integration by parts will just give that

$$\int e^{2x} \cos(3x) dx = \int e^{2x} \cos(3x) dx$$

which does not help in solving the problem. With the correct choice of u and dv , $u = \frac{2}{3}e^{2x}$ and $dv = \sin(3x) dx$, we have that $du = \frac{4}{3}e^{2x} dx$ and $v = -\frac{1}{3}\cos(3x)$, so that integration by parts tells us that

$$\int \frac{2}{3}e^{2x} \sin(3x) dx = -\frac{2}{9}e^{2x} \cos(3x) - \int -\frac{4}{9}e^{2x} \cos(3x) dx.$$

Combining this with our first integration by parts gives

$$\begin{aligned} \int e^{2x} \cos(3x) dx &= \frac{1}{3}e^{2x} \sin(3x) - \int \frac{2}{3}e^{2x} \sin(3x) dx \\ &= \frac{1}{3}e^{2x} \sin(3x) + \frac{2}{9}e^{2x} \cos(3x) - \int \frac{4}{9}e^{2x} \cos(3x) dx \\ \int e^{2x} \cos(3x) dx &= \frac{1}{3}e^{2x} \sin(3x) + \frac{2}{9}e^{2x} \cos(3x) - \frac{4}{9} \int e^{2x} \cos(3x) dx. \end{aligned}$$

In this case, we can see that the integral on the left matches the integral on the right. If we combine these on the left side, we get

$$\frac{13}{9} \int e^{2x} \cos(3x) dx = \frac{1}{3}e^{2x} \sin(3x) + \frac{2}{9}e^{2x} \cos(3x)$$

which then allows us to solve for the answers as

$$\int e^{2x} \cos(3x) dx = \frac{3}{13}e^{2x} \sin(3x) + \frac{2}{13}e^{2x} \cos(3x).$$

\boxed{}

Appendix C

Table of Laplace Transforms

The function u is the Heaviside function, δ is the Dirac delta function, and

$$\Gamma(t) = \int_0^\infty e^{-\tau} \tau^{t-1} d\tau, \quad \text{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-\tau^2} d\tau, \quad \text{erfc}(t) = 1 - \text{erf}(t).$$

$f(t)$	$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty e^{-st} f(t) dt$
C	$\frac{C}{s}$
t	$\frac{1}{s^2}$
t^2	$\frac{2}{s^3}$
t^n	$\frac{n!}{s^{n+1}}$
$t^p \quad (p > 0)$	$\frac{\Gamma(p+1)}{s^{p+1}}$
e^{-at}	$\frac{1}{s+a}$
$\sin(\omega t)$	$\frac{\omega}{s^2 + \omega^2}$
$\cos(\omega t)$	$\frac{s}{s^2 + \omega^2}$
$\sinh(\omega t)$	$\frac{\omega}{s^2 - \omega^2}$
$\cosh(\omega t)$	$\frac{s}{s^2 - \omega^2}$
$u(t - a)$	$\frac{e^{-as}}{s}$
$\delta(t)$	1
$\delta(t - a)$	e^{-as}
$\text{erf}\left(\frac{t}{2a}\right)$	$\frac{1}{s} e^{(as)^2} \text{erfc}(as)$
$\frac{1}{\sqrt{\pi t}} \exp\left(\frac{-a^2}{4t}\right) \quad (a \geq 0)$	$\frac{e^{-as}}{\sqrt{s}}$
$\frac{1}{\sqrt{\pi t}} - ae^{a^2 t} \text{erfc}(a\sqrt{t}) \quad (a > 0)$	$\frac{1}{\sqrt{s+a}}$

$f(t)$	$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty e^{-st} f(t) dt$
$af(t) + bg(t)$	$aF(s) + bG(s)$
$f(at) \quad (a > 0)$	$\frac{1}{a}F\left(\frac{s}{a}\right)$
$f(t - a)u(t - a)$	$e^{-as}F(s)$
$e^{-at}f(t)$	$F(s + a)$
$g'(t)$	$sG(s) - g(0)$
$g''(t)$	$s^2G(s) - sg(0) - g'(0)$
$g'''(t)$	$s^3G(s) - s^2g(0) - sg'(0) - g''(0)$
$g^{(n)}(t)$	$s^nG(s) - s^{n-1}g(0) - \dots - g^{(n-1)}(0)$
$(f * g)(t) = \int_0^t f(\tau)g(t - \tau) d\tau$	$F(s)G(s)$
$tf(t)$	$-F'(s)$
$t^n f(t)$	$(-1)^n F^{(n)}(s)$
$\int_0^t f(\tau) d\tau$	$\frac{1}{s}F(s)$
$\frac{f(t)}{t}$	$\int_s^\infty F(\sigma) d\sigma$

Further Reading

- [BM] Paul W. Berg and James L. McGregor, *Elementary Partial Differential Equations*, Holden-Day, San Francisco, CA, 1966.
- [BD] William E. Boyce and Richard C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, 11th edition, John Wiley & Sons Inc., New York, NY, 2017.
- [EP] C.H. Edwards and D.E. Penney, *Differential Equations and Boundary Value Problems: Computing and Modeling*, 5th edition, Pearson, 2014.
- [F] Stanley J. Farlow, *An Introduction to Differential Equations and Their Applications*, McGraw-Hill, Inc., Princeton, NJ, 1994. (Published also by Dover Publications, 2006.)
- [I] E.L. Ince, *Ordinary Differential Equations*, Dover Publications, Inc., New York, NY, 1956.
- [JL] Jiří Lebl, *Notes on Diffy Qs*, Open-source publication, <https://www.jirka.org/diffyqs>.
- [SZ] Carl Stitz and Jeff Zeager, *Precalculus*. Version 4. 2017. <https://www.stitz-zeager.com/>
- [T] William F. Trench, *Elementary Differential Equations with Boundary Value Problems*. Books and Monographs. Book 9. 2013. <https://digitalcommons.trinity.edu/mono/9>
- [ZW] Dennis Zill and Warren Wright, *Advanced Engineering Mathematics*, 6th Edition, Jones & Bartlett Learning, 2016.

Answers to Selected Exercises

0.1.5: Compute $x' = -2e^{-2t}$ and $x'' = 4e^{-2t}$. Then $(4e^{-2t}) + 4(-2e^{-2t}) + 4(e^{-2t}) = 0$.

0.1.8: Yes.

0.1.10: $y = x^r$ is a solution for $r = 0$ and $r = 2$.

0.1.13: $C_1 = 100, C_2 = -90$

0.1.15: $\varphi = -9e^{8s}$

0.1.17: a) $x = 9e^{-4t}$ b) $x = \cos(2t) + \sin(2t)$ c) $p = 4e^{3q}$ d) $T = 3 \sinh(2x)$

0.2.2: a) PDE, equation, second order, linear, nonhomogeneous, constant coefficient.

b) ODE, equation, first order, linear, nonhomogeneous, not constant coefficient, not autonomous.

c) ODE, equation, seventh order, linear, homogeneous, constant coefficient, autonomous.

d) ODE, equation, second order, linear, nonhomogeneous, constant coefficient, autonomous.

e) ODE, system, second order, nonlinear.

f) PDE, equation, second order, nonlinear.

0.2.6: equation: $a(x)y = b(x)$, solution: $y = \frac{b(x)}{a(x)}$.

0.2.7: $k = 0$ or $k = 1$

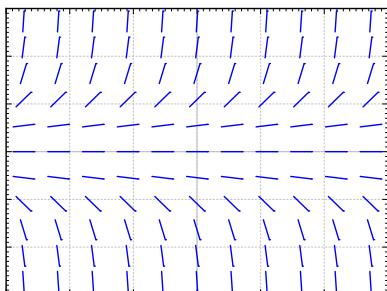
1.1.4: $y = e^x + \frac{x^2}{2} + 9$

1.1.11: 170

1.1.15: The equation is $r' = -C$ for some constant C . The snowball will be completely melted in 25 minutes from time $t = 0$.

1.1.16: $y = Ax^3 + Bx^2 + Cx + D$, so 4 constants.

1.2.2:



$y = 0$ is a solution such that $y(0) = 0$.

1.2.7: a) $y' = \cos y$, b) $y' = y \cos(x)$, c) $y' = \sin x$. Justification left to reader.

1.3.2: $x = (3t - 2)^{1/3}$

1.3.4: $x = \sin^{-1}(t + 1)$

1.3.8: a) $\frac{y^2}{2} = x^2 + C$ b) $y = 2\sqrt{x^2 + 3}$ c) $y = -2\sqrt{x^2 + 1}$

1.3.9: If $n \neq 1$, then $y = ((1-n)x + 1)^{1/(1-n)}$. If $n = 1$, then $y = e^x$.

1.3.12: $y = Ce^{x^2}$

1.3.14: $x = e^{t^3} + 1$

1.3.17: $x^3 + x = t + 2$

1.3.20: $\sin(y) = -\cos(x) + C$

1.3.23: $y = \frac{1}{1-\ln x}$

1.3.29: The range is approximately 7.45 to 12.15 minutes.

1.3.30: a) $x = \frac{1000e^t}{e^t+24}$. b) 102 rabbits after one month, 861 after 5 months, 999 after 10 months, 1000 after 15 months.

1.4.13: $y = Ce^{-x^3} + 1/3$

1.4.17: $y = 2e^{\cos(2x)+1} + 1$

1.5.14: Yes a solution exists. $y' = f(x, y)$ where $f(x, y) = xy$. The function $f(x, y)$ is continuous and $\frac{\partial f}{\partial y} = x$, which is also continuous near $(0, 0)$. So a solution exists and is unique. (In fact $y = 0$ is the solution).

1.5.15: No, the equation is not defined at $(x, y) = (1, 0)$.

1.5.16: Picard does not apply as f is not continuous at $y = 0$. The equation does not have a continuously differentiable solution. Suppose it did. Notice that $y'(0) = 1$. By the first derivative test, $y(x) > 0$ for small positive x . But then for those x we would have $y'(x) = 0$, so clearly the derivative cannot be continuous.

1.5.17: The solution is $y(x) = \int_{x_0}^x f(s) ds + y_0$, and this does indeed exist for every x .

1.6.7: Approximately: 1.0000, 1.2397, 1.3829

1.6.9: a) 0, 8, 12 b) $x(4) = 16$, so errors are: 16, 8, 4. c) Factors are 0.5, 0.5, 0.5.

1.6.10: a) 0, 0, 0 b) $x = 0$ is a solution so errors are: 0, 0, 0.

1.6.12: a) Improved Euler: $y(1) \approx 3.3897$ for $h = 1/4$, $y(1) \approx 3.4237$ for $h = 1/8$, b) Standard Euler: $y(1) \approx 2.8828$ for $h = 1/4$, $y(1) \approx 3.1316$ for $h = 1/8$, c) $y = 2e^x - x - 1$, so $y(2)$ is approximately 3.4366. d) Approximate errors for improved Euler: 0.046852 for $h = 1/4$, and 0.012881 for $h = 1/8$. For standard Euler: 0.55375 for $h = 1/4$, and 0.30499 for $h = 1/8$. Factor is approximately 0.27 for improved Euler, and 0.55 for standard Euler.

1.7.5: a) 0, 1, 2 are critical points. b) $x = 0$ is unstable (semistable), $x = 1$ is asymptotically stable, and $x = 2$ is unstable. c) 1

1.7.8: a) There are no critical points. b) ∞

1.7.10: a) α is a stable critical point, β is an unstable one. b) α , c) α , d) ∞ or DNE.

1.7.12: a) $\frac{dx}{dt} = kx(M - x) + A$ b) $\frac{kM + \sqrt{(kM)^2 + 4Ak}}{2k}$

1.8.3: a) $e^{xy} + \sin(x) = C$ b) $x^2 + xy - 2y^2 = C$ c) $e^x + e^y = C$ d) $x^3 + 3xy + y^3 = C$

1.8.10: a) Integrating factor is y , equation becomes $dx + 3y^2 dy = 0$. b) Integrating factor is e^x , equation becomes $e^x dx - e^{-y} dy = 0$. c) Integrating factor is y^2 , equation becomes $(\cos(x) + y) dx + x dy = 0$. d) Integrating factor is x , equation becomes $(2xy + y^2) dx + (x^2 + 2xy) dy = 0$.

1.8.15: a) The equation is $-f(x) dx + \frac{1}{g(y)} dy$, and this is exact because $M = -f(x)$, $N = \frac{1}{g(y)}$, so $M_y = 0 = N_x$. b) $-x dx + \frac{1}{y} dy = 0$, leads to potential function $F(x, y) = -\frac{x^2}{2} + \ln|y|$, solving $F(x, y) = C$ leads to the same solution as the example.

1.9.5: 250 grams

1.9.6: $P(5) = 1000e^{2 \times 5 - 0.05 \times 5^2} = 1000e^{8.75} \approx 6.31 \times 10^6$

1.9.7: $Ah' = I - kh$, where k is a constant with units m^2/s .

1.10.2: $y = \frac{2}{3x-2}$

1.10.4: $y = \frac{3-x^2}{2x}$

1.10.9: $y = (7e^{3x} + 3x + 1)^{1/3}$

1.10.13: $y = \sqrt{x^2 - \ln(C-x)}$

2.1.5: Yes. To justify try to find a constant A such that $\sin(x) = Ae^x$ for all x .

2.1.6: No. $e^{x+2} = e^2 e^x$.

2.1.7: $y = 5$

2.1.13: $y = C_1 \ln(x) + C_2$

2.1.21: $y = C_1 e^{(-2+\sqrt{2})x} + C_2 e^{(-2-\sqrt{2})x}$

2.1.22: $y = \frac{2(a-b)}{5} e^{-3x/2} + \frac{3a+2b}{5} e^x$

2.1.23: $y = \frac{a\beta-b}{\beta-\alpha} e^{\alpha x} + \frac{b-a\alpha}{\beta-\alpha} e^{\beta x}$

2.1.24: $y'' - y' - 6y = 0$

2.1.25: $y'' - 3y' + 2y = 0$

2.2.5: $3\sqrt{2} \cos(2x - \frac{\pi}{4})$

2.2.13: $y = e^{-x/4} \cos((\sqrt{7}/4)x) - \sqrt{7}e^{-x/4} \sin((\sqrt{7}/4)x)$

2.2.14: $z(t) = 2e^{-t} \cos(t)$

2.3.4: $y = C_1 e^{3x} + C_2 x e^{3x}$

2.3.9: c) $y(x) = C_1 x + C_2 \frac{1}{x^3}$

2.3.10: c) $y(x) = C_1 \frac{1}{x} + C_2 \frac{1}{x^2}$

2.3.11: c) $y(x) = C_1 x^2 + C_2 x^5$

2.4.5: $k = 8/9$ (and larger)

2.4.8: a) $k = 500000$ b) $\frac{1}{5\sqrt{2}} \approx 0.141$ c) 45000 kg d) 11250 kg

2.4.10: $m_0 = \frac{1}{3}$. If $m < m_0$, then the system is overdamped and will not oscillate.

2.4.11: a) $0.05I'' + 0.1I' + (\frac{1}{5})I = 0$ b) $I = Ce^{-t} \cos(\sqrt{3}t - \gamma)$ c) $I = 10e^{-t} \cos(\sqrt{3}t) + \frac{10}{\sqrt{3}}e^{-t} \sin(\sqrt{3}t)$

2.5.4: $y = \frac{-16\sin(3x)+6\cos(3x)}{73}$

2.5.8: $y(x) = x^2 - 4x + 6 + e^{-x}(x - 5)$

2.5.10: a) $y = \frac{2e^x+3x^3-9x}{6}$ b) $y = C_1 \cos(\sqrt{2}x) + C_2 \sin(\sqrt{2}x) + \frac{2e^x+3x^3-9x}{6}$

2.5.19: $y = \frac{2xe^x-(e^x+e^{-x})\log(e^{2x}+1)}{4}$

2.5.20: $y = \frac{-\sin(x+c)}{3} + C_1 e^{\sqrt{2}x} + C_2 e^{-\sqrt{2}x}$

2.6.6: $x_{sp} = \frac{(\omega_0^2 - \omega^2)F_0}{m(2\omega p)^2 + m(\omega_0^2 - \omega^2)^2} \cos(\omega t) + \frac{2\omega p F_0}{m(2\omega p)^2 + m(\omega_0^2 - \omega^2)^2} \sin(\omega t) + \frac{A}{k}$, where $p = \frac{c}{2m}$ and $\omega_0 = \sqrt{\frac{k}{m}}$.

2.6.9: $\omega = \frac{\sqrt{31}}{4\sqrt{2}} \approx 0.984$ $C(\omega) = \frac{16}{3\sqrt{7}} \approx 2.016$

2.6.12: a) $\omega = 2$ b) 25

2.7.2: $y = C_1 e^x + C_2 x^3 + C_3 x^2 + C_4 x + C_5$

2.7.7: a) $r^3 - 3r^2 + 4r - 12 = 0$ b) $y''' - 3y'' + 4y' - 12y = 0$ c) $y = C_1 e^{3x} + C_2 \sin(2x) + C_3 \cos(2x)$

2.7.9: $y(x) = C_1 e^{4x} + C_2 e^{-x} + C_3 e^{-x} \cos(2x) + C_4 e^{-x} \sin(2x)$

2.7.16: No. $e^1 e^x - e^{x+1} = 0$.

2.7.19: Yes. (Hint: First note that $\sin(x)$ is bounded. Then note that x and $x \sin(x)$ cannot be multiples of each other.)

2.7.20: $y = 0$

2.7.22: $y''' - y'' + y' - y = 0$

3.1.5: a) $\sqrt{10}$ b) $\sqrt{14}$ c) 3

3.1.7: a) $\begin{bmatrix} 9 \\ -2 \end{bmatrix}$ b) $\begin{bmatrix} -3 \\ 3 \end{bmatrix}$ c) $\begin{bmatrix} 5 \\ -3 \end{bmatrix}$ d) $\begin{bmatrix} -4 \\ 8 \end{bmatrix}$ e) $\begin{bmatrix} 3 \\ 7 \end{bmatrix}$ f) $\begin{bmatrix} -8 \\ 3 \end{bmatrix}$

3.1.9: a) $\begin{bmatrix} \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ b) $\begin{bmatrix} \frac{1}{\sqrt{6}} \\ \frac{-1}{\sqrt{6}} \end{bmatrix}$ c) $\left(\frac{2}{\sqrt{33}}, \frac{-5}{\sqrt{33}}, \frac{2}{\sqrt{33}} \right)$

3.1.14: a) 20 b) 10 c) 20

3.1.18: a) $(3, -1)$ b) $(4, 0)$ c) $(-1, -1)$

3.2.2: a) $\begin{bmatrix} 7 & 4 & 4 \\ 2 & 3 & 4 \end{bmatrix}$ b) $\begin{bmatrix} 5 & -3 & 0 \\ 13 & 10 & 6 \\ -1 & 3 & 1 \end{bmatrix}$

3.2.4: a) $\begin{bmatrix} -1 & 13 \\ 9 & 14 \end{bmatrix}$ b) $\begin{bmatrix} 2 & -5 \\ 5 & 5 \end{bmatrix}$

3.2.6: a) $\begin{bmatrix} 22 & 31 \\ 42 & 44 \end{bmatrix}$ b) $\begin{bmatrix} 18 & 18 & 12 \\ 6 & 0 & 8 \\ 34 & 48 & -2 \end{bmatrix}$ c) $\begin{bmatrix} 11 & 12 & 36 & 14 \\ -2 & 4 & 5 & -2 \\ 13 & 38 & 20 & 28 \end{bmatrix}$ d) $\begin{bmatrix} -2 & -12 \\ 3 & 24 \\ 1 & 9 \end{bmatrix}$

3.2.9: a) $\begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$ b) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ c) $\begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix}$ d) $\begin{bmatrix} 1/2 & -1/4 \\ -1/2 & 1/2 \end{bmatrix}$

3.2.11: a) $\begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \end{bmatrix}$ b) $\begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/5 & 0 \\ 0 & 0 & -1 \end{bmatrix}$ c) $\begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix}$

3.3.2: a) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ b) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ c) $\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$ d) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1/3 \\ 0 & 0 & 0 \end{bmatrix}$ e) $\begin{bmatrix} 1 & 0 & 0 & 77/15 \\ 0 & 1 & 0 & -2/15 \\ 0 & 0 & 1 & -8/5 \end{bmatrix}$

f) $\begin{bmatrix} 1 & 0 & -1/2 & 0 \\ 0 & 1 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ g) $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ h) $\begin{bmatrix} 1 & 2 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

3.3.4: a) $\begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ b) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & -1 & 0 \end{bmatrix}$ c) $\begin{bmatrix} 5/2 & 1 & -3 \\ -1 & -1/2 & 3/2 \\ -1 & 0 & 1 \end{bmatrix}$

3.3.6: a) $x_1 = -2, x_2 = 7/3$ b) no solution c) $a = -3, b = 10, c = -8$ d) x_3 is free, $x_1 = -1 + 3x_3, x_2 = 2 - x_3$

3.3.8: a) $\begin{bmatrix} -1 \\ 3 \end{bmatrix}$ b) $\begin{bmatrix} -3 \\ 1 \end{bmatrix}$

3.4.2: a) 3 b) 1 c) 2

3.4.5: a) $[1 \ 0 \ 0], [0 \ 1 \ 0], [0 \ 0 \ 1]$ b) $[1 \ 1 \ 1]$ c) $[1 \ 0 \ 1/3], [0 \ 1 \ -1/3]$

3.4.6: a) $\begin{bmatrix} 7 \\ 7 \\ 7 \end{bmatrix}, \begin{bmatrix} -1 \\ 7 \\ 6 \end{bmatrix}, \begin{bmatrix} 7 \\ 6 \\ 2 \end{bmatrix}$ b) $\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$ c) $\begin{bmatrix} 0 \\ 6 \\ 4 \end{bmatrix}, \begin{bmatrix} 3 \\ 3 \\ 7 \end{bmatrix}$

3.4.8: $\begin{bmatrix} 3 \\ 1 \\ -5 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \\ -1 \end{bmatrix}$

3.4.10: a) $\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ dimension 2, b) $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$ dimension 2, c) $\begin{bmatrix} 5 \\ 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 5 \\ -1 \\ 5 \end{bmatrix}, \begin{bmatrix} -1 \\ 3 \\ -4 \end{bmatrix}$

dimension 3, d) $\begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}$ dimension 2, e) $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ dimension 1, f) $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$ dimension 2

3.5.3: a) -2 b) 8 c) 0 d) -6 e) -3 f) 28 g) 16 h) -24

3.5.5: a) 3 b) 9 c) 3 d) $1/4$

3.5.7: Rank is 3. Therefore A is not invertible (since the rank is not 4), and there are non-zero solutions to $A\vec{x} = \vec{0}$.

3.5.8: Rank is 3. Therefore A is invertible, and there is exactly one solution to $A\vec{x} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$,

namely $A^{-1} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

3.5.9: -1**3.5.10:** 2**3.5.11:** 8**3.5.13:** 1/12**3.5.16:** 1 and 3

$$\mathbf{3.6.1:} \quad \lambda_1 = -2, \vec{v}_1 = \begin{bmatrix} -3 \\ 1 \end{bmatrix}, \lambda_2 = 4, \vec{v}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}.$$

$$\mathbf{3.6.2:} \quad \lambda_1 = -2, \vec{v}_1 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \lambda_2 = -4, \vec{v}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

$$\mathbf{3.6.3:} \quad \lambda_1 = -4, \vec{v}_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \lambda_2 = -3, \vec{v}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}.$$

$$\mathbf{3.6.4:} \quad \lambda_1 = 3 + 2i, \vec{v}_1 = \begin{bmatrix} 3 - i \\ 4 \end{bmatrix}, \lambda_2 = 3 - 2i, \vec{v}_2 = \begin{bmatrix} 3 + i \\ 4 \end{bmatrix}.$$

$$\mathbf{3.6.5:} \quad \lambda_1 = -1 + i, \vec{v}_1 = \begin{bmatrix} 2 \\ -1 + i \end{bmatrix}, \lambda_2 = -1 - i, \vec{v}_2 = \begin{bmatrix} 2 \\ -1 - i \end{bmatrix}.$$

$$\mathbf{3.6.6:} \quad \lambda_1 = -2 + 2i, \vec{v}_1 = \begin{bmatrix} 1 - i \\ 4 \end{bmatrix}, \lambda_2 = -2 - 2i, \vec{v}_2 = \begin{bmatrix} 1 + i \\ 4 \end{bmatrix}.$$

$$\mathbf{3.6.7:} \quad \lambda_1 = 4, \vec{v}_1 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

$$\mathbf{3.6.8:} \quad \lambda_1 = -3, \vec{v}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$\mathbf{3.6.9:} \quad \lambda_1 = 2, \vec{v}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \lambda_2 = 1, \vec{v}_2 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, \lambda_3 = 4, \vec{v}_3 = \begin{bmatrix} 1 \\ -3 \\ -2 \end{bmatrix}$$

$$\mathbf{3.6.10:} \quad \lambda_1 = -4, \vec{v}_1 = \begin{bmatrix} 1 \\ 3 \\ -3 \end{bmatrix}, \lambda_2 = -3, \vec{v}_2 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, \lambda_3 = -1, \vec{v}_3 = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}$$

$$\mathbf{3.6.11:} \quad \lambda_1 = 1 + 3i, \vec{v}_1 = \begin{bmatrix} 0 \\ 2 \\ -1 + i \end{bmatrix}, \lambda_2 = 1 - 3i, \vec{v}_2 = \begin{bmatrix} 0 \\ 2 \\ -1 - i \end{bmatrix}, \lambda_3 = -2, \vec{v}_3 = \begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix}$$

$$\mathbf{3.6.12:} \quad \lambda_1 = 2, \vec{v}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \lambda_2 = 1, \vec{v}_2 = \begin{bmatrix} 0 \\ -2 \\ 1 \end{bmatrix} \text{ (double root)}$$

$$\mathbf{3.6.13:} \quad \lambda_1 = -2, \vec{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \lambda_2 = 1, \vec{v}_2 = \begin{bmatrix} 3 \\ -4 \end{bmatrix}.$$

$$\mathbf{3.6.14:} \quad \lambda_1 = 2, \vec{v}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \text{ Generalized eigenvector } \vec{w} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

$$\mathbf{3.6.18:} \quad \begin{bmatrix} 3 & 0 & 0 \\ 0 & -2 & 1 \\ 0 & 0 & -2 \end{bmatrix}$$

3.7.2: a) $\begin{bmatrix} 3 \\ -1 \\ 0 \\ 0 \end{bmatrix}$, b) $\begin{bmatrix} 3 \\ 0 \\ 3 \\ -1 \end{bmatrix}$ c) $\begin{bmatrix} -1 \\ -1 \\ 0 \end{bmatrix}$ d) $\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$, e) $\begin{bmatrix} -1 \\ 0 \\ 0 \\ -1 \end{bmatrix}$

3.7.6: a) 3 b) 2 c) 3 d) 2 e) 3

4.1.6: $y_1 = C_1 e^{3x}$, $y_2 = y(x) = C_2 e^x + \frac{C_1}{2} e^{3x}$, $y_3 = y(x) = C_3 e^x + \frac{C_1}{2} e^{3x}$

4.1.7: $x = \frac{5}{3}e^{2t} - \frac{2}{3}e^{-t}$, $y = \frac{5}{3}e^{2t} + \frac{4}{3}e^{-t}$

4.1.10: $x'_1 = x_2$, $x'_2 = x_3$, $x'_3 = x_1 + t$

4.1.11: $y'_3 + y_1 + y_2 = t$, $y'_4 + y_1 - y_2 = t^2$, $y'_1 = y_3$, $y'_2 = y_4$

4.1.17: $x_1 = x_2 = at$. Explanation of the intuition is left to reader.

4.1.19: a) Left to reader. b) $x'_1 = \frac{r}{V}(x_2 - x_1)$, $x'_2 = \frac{r}{V}x_1 - \frac{r-s}{V}x_2$. c) As t goes to infinity, both x_1 and x_2 go to zero, explanation is left to reader.

4.2.7: -15

4.2.11: -2

4.2.15: $\vec{x} = \begin{bmatrix} 15 \\ -5 \end{bmatrix}$

4.2.18: a) $\begin{bmatrix} 1/a & 0 \\ 0 & 1/b \end{bmatrix}$ b) $\begin{bmatrix} 1/a & 0 & 0 \\ 0 & 1/b & 0 \\ 0 & 0 & 1/c \end{bmatrix}$

4.2.20: $\lambda_1 = 1$, $\vec{v}_1 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$, $\lambda_2 = 2$, $\vec{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

4.2.21: $\lambda_1 = -2 + 2i$, $\vec{v}_1 = \begin{bmatrix} -3+i \\ 4 \end{bmatrix}$, $\lambda_2 = -2 - 2i$, $\vec{v}_2 = \begin{bmatrix} -3-i \\ 4 \end{bmatrix}$.

4.2.22: $\lambda_1 = 4$, $\vec{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$, $\lambda_2 = -2$, $\vec{v}_2 = \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix}$, $\lambda_3 = -3$, $\vec{v}_3 = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$.

4.3.2: $\begin{bmatrix} x \\ y \end{bmatrix}' = \begin{bmatrix} 3 & -1 \\ t & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e^t \\ 0 \end{bmatrix}$

4.3.5: Yes.

4.3.7: No. $2 \begin{bmatrix} \cosh(t) \\ 1 \end{bmatrix} - \begin{bmatrix} e^t \\ 1 \end{bmatrix} - \begin{bmatrix} e^{-t} \\ 1 \end{bmatrix} = \vec{0}$

4.3.10: a) $\vec{x}' = \begin{bmatrix} 0 & 2t \\ 0 & 2t \end{bmatrix} \vec{x}$ b) $\vec{x} = \begin{bmatrix} C_2 e^{t^2} + C_1 \\ C_2 e^{t^2} \end{bmatrix}$

4.4.6: $\vec{x} = C_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^t + C_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-t}$

4.4.8: $\vec{x} = C_1 \begin{bmatrix} \cos(t) \\ -\sin(t) \end{bmatrix} + C_2 \begin{bmatrix} \sin(t) \\ \cos(t) \end{bmatrix}$

4.4.10: a) Eigenvalues: 4, 0, -1 Eigenvectors: $\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 3 \\ 5 \\ -2 \end{bmatrix}$

b) $\vec{x} = C_1 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} e^{4t} + C_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + C_3 \begin{bmatrix} 3 \\ 5 \\ -2 \end{bmatrix} e^{-t}$

4.4.14: a) Eigenvalues: $\frac{1+\sqrt{3}i}{2}, \frac{1-\sqrt{3}i}{2}$, Eigenvectors: $\begin{bmatrix} -2 \\ 1-\sqrt{3}i \\ 1+\sqrt{3}i \end{bmatrix}$, $\begin{bmatrix} -2 \\ 1+\sqrt{3}i \\ 1-\sqrt{3}i \end{bmatrix}$

b) $\vec{x} = C_1 e^{t/2} \begin{bmatrix} -2 \cos\left(\frac{\sqrt{3}t}{2}\right) \\ \cos\left(\frac{\sqrt{3}t}{2}\right) + \sqrt{3} \sin\left(\frac{\sqrt{3}t}{2}\right) \end{bmatrix} + C_2 e^{t/2} \begin{bmatrix} -2 \sin\left(\frac{\sqrt{3}t}{2}\right) \\ \sin\left(\frac{\sqrt{3}t}{2}\right) - \sqrt{3} \cos\left(\frac{\sqrt{3}t}{2}\right) \end{bmatrix}$

4.4.16: a) 3, 0, 0 b) No defects. c) $\vec{x} = C_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{3t} + C_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + C_3 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

4.4.18: a) 1, 1, 2

b) Eigenvalue 1 has a defect of 1

c) $\vec{x} = C_1 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^t + C_2 \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right) e^t + C_3 \begin{bmatrix} 3 \\ -2 \end{bmatrix} e^{2t}$

4.4.20: a) 2, 2, 2

b) Eigenvalue 2 has a defect of 2

c) $\vec{x} = C_1 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{2t} + C_2 \left(\begin{bmatrix} 0 \\ 1 \end{bmatrix} + t \begin{bmatrix} 0 \\ 3 \end{bmatrix} \right) e^{2t} + C_3 \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \frac{t^2}{2} \begin{bmatrix} 0 \\ 3 \end{bmatrix} \right) e^{2t}$

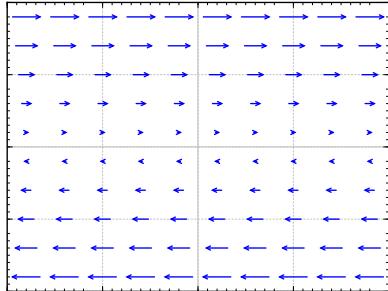
4.4.29: $A = \begin{bmatrix} 5 & 5 \\ 0 & 5 \end{bmatrix}$

- 4.5.4:** a) Two eigenvalues: $\pm\sqrt{2}$ so the behavior is a saddle. b) Two eigenvalues: 1 and 2, so the behavior is a source. c) Two eigenvalues: $\pm 2i$, so the behavior is a center (ellipses). d) Two eigenvalues: -1 and -2, so the behavior is a sink. e) Two eigenvalues: 5 and -3, so the behavior is a saddle.

4.5.6: Spiral source.

- 4.5.7:** a) Nodal source c) Spiral source c) Saddle c) Nodal sink e) Spiral sink f) Improper nodal sink

4.5.9:



The solution does not move anywhere if $y = 0$. When y is positive, the solution moves (with constant speed) in the positive x direction. When y is negative, the solution moves (with constant speed) in the negative x direction. It is not one of the behaviors we have seen.

Note that the matrix has a double eigenvalue 0 and the general solution is $x = C_1 t + C_2$ and $y = C_1$, which agrees with the description above.

4.6.5: The general solution is (particular solutions should agree with one of these):

$$x(t) = C_1 e^{9t} + 4C_2 e^{4t} - \frac{t}{3} - \frac{5}{54}, \quad y(t) = C_1 e^{9t} - C_2 e^{4t} + \frac{t}{6} + \frac{7}{216}$$

4.6.7: The general solution is (particular solutions should agree with one of these):

$$x(t) = C_1 e^t + C_2 e^{-t} + te^t, \quad y(t) = C_1 e^t - C_2 e^{-t} + te^t$$

4.6.8: $\vec{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(\frac{5}{2} e^t - t - 1 \right) + \begin{bmatrix} 1 \\ -1 \end{bmatrix} \frac{-1}{2} e^{-t}$

4.7.4: $\vec{x} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} (a_1 \cos(\sqrt{3}t) + b_1 \sin(\sqrt{3}t)) + \begin{bmatrix} 0 \\ 1 \\ -2 \end{bmatrix} (a_2 \cos(\sqrt{2}t) + b_2 \sin(\sqrt{2}t)) + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} (a_3 \cos(t) + b_3 \sin(t)) + \begin{bmatrix} -1 \\ 1/2 \\ 2/3 \end{bmatrix} \cos(2t)$

4.7.8: $\begin{bmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & m \end{bmatrix} \vec{x}'' = \begin{bmatrix} -k & k & 0 \\ k & -2k & k \\ 0 & k & -k \end{bmatrix} \vec{x}$. Solution: $\vec{x} = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} (a_1 \cos(\sqrt{3k/m}t) + b_1 \sin(\sqrt{3k/m}t)) + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} (a_2 \cos(\sqrt{k/m}t) + b_2 \sin(\sqrt{k/m}t)) + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} (a_3 t + b_3)$.

4.7.9: $x_2 = (2/5) \cos(\sqrt{1/6}t) - (2/5) \cos(t)$

4.7.12: $\vec{x} = [\begin{smallmatrix} 1 \\ 9 \end{smallmatrix}] \left(\left(\frac{1}{140} + \frac{1}{120\sqrt{6}} \right) e^{\sqrt{6}t} + \left(\frac{1}{140} + \frac{1}{120\sqrt{6}} \right) e^{-\sqrt{6}t} - \frac{t}{60} - \frac{\cos(t)}{70} \right)$
 $+ [\begin{smallmatrix} -9 \\ -1 \end{smallmatrix}] \left(\frac{-9}{80} \sin(2t) + \frac{1}{30} \cos(2t) + \frac{9t}{40} - \frac{\cos(t)}{30} \right)$

4.8.4: $e^{tA} = \begin{bmatrix} \frac{e^{3t}+e^{-t}}{2} & \frac{e^{-t}-e^{3t}}{2} \\ \frac{e^{-t}-e^{3t}}{2} & \frac{e^{3t}+e^{-t}}{2} \end{bmatrix}$

4.8.5: $e^{tA} = \begin{bmatrix} 2e^{3t}-4e^{2t}+3e^t & \frac{3e^t}{2}-\frac{3e^{3t}}{2} & -e^{3t}+4e^{2t}-3e^t \\ 2e^t-2e^{2t} & e^t & 2e^{2t}-2e^t \\ 2e^{3t}-5e^{2t}+3e^t & \frac{3e^t}{2}-\frac{3e^{3t}}{2} & -e^{3t}+5e^{2t}-3e^t \end{bmatrix}$

4.8.6: a) $e^{tA} = \begin{bmatrix} (t+1)e^{2t} & -te^{2t} \\ te^{2t} & (1-t)e^{2t} \end{bmatrix}$ b) $\vec{x} = \begin{bmatrix} (1-t)e^{2t} \\ (2-t)e^{2t} \end{bmatrix}$

4.8.15: $\begin{bmatrix} 1+2t+5t^2 & 3t+6t^2 \\ 2t+4t^2 & 1+2t+5t^2 \end{bmatrix} \quad e^{0.1A} \approx \begin{bmatrix} 1.25 & 0.36 \\ 0.24 & 1.25 \end{bmatrix}$

4.8.17: a) $\begin{bmatrix} 5(3^n) - 2^{n+2} & 4(3^n) - 2^{n+2} \\ 5(2^n) - 5(3^n) & 5(2^n) - 4(3^n) \end{bmatrix}$ b) $\begin{bmatrix} 3 - 2(3^n) & 2(3^n) - 2 \\ 3 - 3^{n+1} & 3^{n+1} - 2 \end{bmatrix}$
c) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ if n is even, and $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ if n is odd.

5.1.3: a) Critical points $(0, 0)$ and $(0, 1)$. At $(0, 0)$ using $u = x, v = y$ the linearization is $u' = -2u - (1/\pi)v, v' = -v$. At $(0, 1)$ using $u = x, v = y - 1$ the linearization is $u' = -2u + (1/\pi)v, v' = v$.

b) Critical point $(0, 0)$. Using $u = x, v = y$ the linearization is $u' = u + v, v' = u$.

c) Critical point $(1/2, -1/4)$. Using $u = x - 1/2, v = y + 1/4$ the linearization is $u' = -u + v, v' = u + v$.

5.1.9: Critical points are $(0, 0, 0)$, and $(-1, 1, -1)$. The linearization at the origin using variables $u = x, v = y, w = z$ is $u' = u, v' = -v, z' = w$. The linearization at the point $(-1, 1, -1)$ using variables $u = x + 1, v = y - 1, w = z + 1$ is $u' = u - 2w, v' = -v - 2w, w' = w - 2u$.

5.1.10: $u' = f(u, v, w), v' = g(u, v, w), w' = 1$.

5.2.2: a) $(0, 0)$: saddle (unstable), $(1, 0)$: source (unstable), b) $(0, 0)$: spiral sink (asymptotically stable), $(0, 1)$: saddle (unstable), c) $(1, 0)$: saddle (unstable), $(0, 1)$: saddle (unstable)

5.2.6: a) $\frac{1}{2}y^2 + \frac{1}{3}x^3 - 4x = C$, critical points: $(-2, 0)$, an unstable saddle, and $(2, 0)$, a stable center. b) $\frac{1}{2}y^2 + e^x = C$, no critical points. c) $\frac{1}{2}y^2 + xe^x = C$, critical point at $(-1, 0)$ is a stable center.

5.2.9: Critical point at $(0, 0)$. Trajectories are $y = \pm\sqrt{2C - (1/2)x^4}$, for $C > 0$, these give closed curves around the origin, so the critical point is a stable center.

5.2.15: A critical point x_0 is stable if $f'(x_0) < 0$ and unstable when $f'(x_0) > 0$.

5.3.2: a) Critical points are $\omega = 0, \theta = k\pi$ for any integer k . When k is odd, we have a saddle point. When k is even we get a sink. b) The findings mean the pendulum will simply go to one of the sinks, for example $(0, 0)$ and it will not swing back and forth. The friction is too high for it to oscillate, just like an overdamped mass-spring system.

5.3.4: a) Solving for the critical points we get $(0, -h/d)$ and $(\frac{bh+ad}{ac}, \frac{a}{b})$. The Jacobian matrix at $(0, -h/d)$ is $\begin{bmatrix} a+bh/d & 0 \\ -ch/d & -d \end{bmatrix}$ whose eigenvalues are $a + bh/d$ and $-d$. So the eigenvalues are always real of opposite signs and we get a saddle (In the application however we are only looking at the positive quadrant so this critical point is not relevant). At $(\frac{bh+ad}{ac}, \frac{a}{b})$ we get Jacobian matrix $\begin{bmatrix} 0 & -\frac{b(bh+ad)}{ac} \\ \frac{ac}{b} & \frac{bh+ad}{a} - d \end{bmatrix}$. b) For the specific numbers given, the second critical point is $(\frac{550}{3}, 40)$ the matrix is $\begin{bmatrix} 0 & -11/6 \\ 3/25 & 1/4 \end{bmatrix}$, which has eigenvalues $\frac{5 \pm i\sqrt{327}}{40}$. Therefore there is a spiral source. This means the solution spirals outwards. The solution will eventually hit one of the axes, $x = 0$ or $y = 0$, so something will die out in the forest.

5.3.5: The critical points are on the line $x = 0$. In the positive quadrant the y' is always positive and so the fox population always grows. The constant of motion is $C = y^a e^{-cx-by}$, for any C this curve must hit the y -axis (why?), so the trajectory will simply approach a point on the y axis somewhere and the number of hares will go to zero.

5.4.3: $(0, 0)$, unstable, $r = \sqrt{3}$, asymptotically stable.

5.4.4: $(0, 0)$, asymptotically stable, $r = \sqrt{2}$, unstable, $r = 2$, asymptotically stable.

5.4.6: Use Bendixson–Dulac Theorem. a) $f_x + g_y = 1 + 1 > 0$, so no closed trajectories. b) $f_x + g_y = -\sin^2(y) + 0 < 0$ for all x, y except the lines given by $y = k\pi$ (where we get zero), so no closed trajectories. c) $f_x + g_y = y + 0 > 0$ for all x, y except the line given by $y = 0$ (where we get zero), so no closed trajectories.

5.4.7: Using Poincaré–Bendixson Theorem, the system has a limit cycle, which is the unit circle centered at the origin as $x = \cos(t) + e^{-t}$, $y = \sin(t) + e^{-t}$ gets closer and closer to the unit circle. Thus we also have that $x = \cos(t)$, $y = \sin(t)$ is the periodic solution.

5.4.11: $f(x, y) = y$, $g(x, y) = \mu(1 - x^2)y - x$. So $f_x + g_y = \mu(1 - x^2)$. The Bendixson–Dulac Theorem says there is no closed trajectory lying entirely in the set $x^2 < 1$.

5.4.13: The closed trajectories are those where $\sin(r) = 0$, therefore, all the circles centered at the origin with radius that is a multiple of π are closed trajectories.

5.5.1: Critical points: $(0, 0, 0)$, $(3\sqrt{8}, 3\sqrt{8}, 27)$, $(-3\sqrt{8}, -3\sqrt{8}, 27)$. Linearization at $(0, 0, 0)$ using $u = x$, $v = y$, $w = z$ is $u' = -10u + 10v$, $v' = 28u - v$, $w' = -(8/3)w$. Linearization at $(3\sqrt{8}, 3\sqrt{8}, 27)$ using $u = x - 3\sqrt{8}$, $v = y - 3\sqrt{8}$, $w = z - 27$ is $u' = -10u + 10v$, $v' = u - v - 3\sqrt{8}w$, $w' = 3\sqrt{8}u + 3\sqrt{8}v - (8/3)w$. Linearization at $(-3\sqrt{8}, -3\sqrt{8}, 27)$ using $u = x + 3\sqrt{8}$, $v = y + 3\sqrt{8}$, $w = z - 27$ is $u' = -10u + 10v$, $v' = u - v + 3\sqrt{8}w$, $w' = -3\sqrt{8}u - 3\sqrt{8}v - (8/3)w$.

$$\mathbf{6.1.7:} \quad \frac{8}{s^3} + \frac{8}{s^2} + \frac{4}{s}$$

$$\mathbf{6.1.12:} \quad 2t^2 - 2t + 1 - e^{-2t}$$

$$\mathbf{6.1.17:} \quad \frac{1}{(s+1)^2}$$

$$\mathbf{6.1.19:} \quad \frac{1}{s^2+2s+2}$$

$$\mathbf{6.2.3:} \quad f(t) = (t-1)(u(t-1) - u(t-2)) + u(t-2)$$

$$\mathbf{6.2.8:} \quad x(t) = (2e^{t-1} - t^2 - 1)u(t-1) - \frac{1}{2}e^{-t} + \frac{3}{2}e^t$$

$$\mathbf{6.2.15:} \quad H(s) = \frac{1}{s+1}$$

6.3.3: $\frac{1}{2}(\cos t + \sin t - e^{-t})$

6.3.8: $\frac{1}{2}(\sin t - t \cos t)$

6.3.10: $\int_0^t f(\tau) (1 - \cos(t - \tau)) d\tau$

6.3.12: $5t - 5 \sin t$

6.4.3: $x(t) = t$

6.4.4: $x(t) = e^{-at}$

6.4.8: $x(t) = (\cos * \sin)(t) = \frac{1}{2}t \sin(t)$

6.4.10: $\delta(t) - \sin(t)$

6.4.11: $3\delta(t - 1) + 2t$

6.5.2: $y = (x - t)u(t - x) + t$

6.5.6: $y = e^{-cx} \sin(t - x)u(t - x)$

6.5.8: $s^2Y(x) - s(1 - x^2) + 3Y''(x) + Y(x) = \frac{x}{s} + \frac{1}{s^2}, \quad Y(-1) = 0, \quad Y(1) = 0.$

6.5.11: $y = tx^2 + \frac{t^3}{3}$

7.1.3: Yes. Radius of convergence is 10.

7.1.6: Yes. Radius of convergence is e .

7.1.9: $\sum_{n=7}^{\infty} \frac{1}{(n-7)!} x^n$

7.1.13: $\frac{1}{1-x} = -\frac{1}{1-(2-x)}$ so $\frac{1}{1-x} = \sum_{n=0}^{\infty} (-1)^{n+1}(x-2)^n$, which converges for $1 < x < 3$.

7.1.16: $f(x) - g(x)$ is a polynomial. Hint: Use Taylor series.

7.2.3: $a_2 = 0, a_3 = 0, a_4 = 0$, recurrence relation (for $k \geq 5$): $a_k = \frac{-2a_{k-5}}{k(k-1)}$, so:

$$y(x) = a_0 + a_1 x - \frac{a_0}{10}x^5 - \frac{a_1}{15}x^6 + \frac{a_0}{450}x^{10} + \frac{a_1}{825}x^{11} - \frac{a_0}{47250}x^{15} - \frac{a_1}{99000}x^{16} + \dots$$

7.2.7: Applying the method of this section directly we obtain $a_k = 0$ for all k and so $y(x) = 0$ is the only solution we find.

7.2.11: a) $a_2 = \frac{1}{2}$, and for $k \geq 1$ we have $a_k = \frac{a_{k-3}+1}{k(k-1)}$, so

$$y(x) = a_0 + a_1 x + \frac{1}{2}x^2 + \frac{a_0+1}{6}x^3 + \frac{a_1+1}{12}x^4 + \frac{3}{40}x^5 + \frac{a_0+2}{30}x^6 + \frac{a_1+2}{42}x^7 + \frac{5}{112}x^8 + \frac{a_0+3}{72}x^9 + \frac{a_1+3}{90}x^{10} + \dots$$

b) $y(x) = \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{12}x^4 + \frac{3}{40}x^5 + \frac{1}{15}x^6 + \frac{1}{21}x^7 + \frac{5}{112}x^8 + \frac{1}{24}x^9 + \frac{1}{30}x^{10} + \dots$

7.3.5: $y = x^{3/2} \sum_{k=0}^{\infty} \frac{(-1)^{-1}}{k!(k+2)!} x^k$ (Note that for convenience we did not pick $a_0 = 1$)

7.3.8: $y = Ax^{\frac{1+\sqrt{5}}{2}} + Bx^{\frac{1-\sqrt{5}}{2}}$

7.3.10: $y = Ax + Bx \ln(x)$

7.3.12: a) ordinary, b) singular but not regular singular, c) regular singular, d) regular singular, e) ordinary.

8.1.2: $s = -2$

8.1.4: $\theta \approx 0.3876$

8.1.6: a) -15 b) -1 c) 28

8.1.8: a) $(-\frac{1}{2}, 0, \frac{1}{2})$ b) $(0, 0, 0)$ c) $(2, 0, -2)$

8.1.10: a) $(1, 1, -1) - (2, -1, 1) + 2(1, -5, 3)$ b) $2(2, -1, 1) + (1, -5, 3)$ c) $2(1, 1, -1) - 2(2, -1, 1) + 2(1, -5, 3)$

8.1.12: $(2, -1, 1), (\frac{2}{3}, \frac{8}{3}, \frac{4}{3})$

8.1.14: $(1, 1, -1), (0, 1, 1), (\frac{4}{3}, -\frac{2}{3}, \frac{2}{3})$

8.2.2: a) No. b) Yes. c) Yes. d) Yes.

8.2.3: $\lambda_1 = 2, \vec{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \lambda_2 = 5, \vec{v}_2 = \begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix}$. Not orthogonal.

8.2.5: $\lambda_1 = 1, \vec{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \lambda_2 = 3, \vec{v}_2 = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}, \lambda_3 = -2, \vec{v}_3 = \begin{bmatrix} 0 \\ 1 \\ -2 \end{bmatrix}$. Orthogonal.

8.2.8: $P = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}, D = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$.

8.2.9: There are complex eigenvalues, so this can not be done.

8.2.10: $P = \begin{bmatrix} -1 & 0 \\ 2 & 0.5 \end{bmatrix}, J = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$.

8.2.11: $P = \begin{bmatrix} 0 & 1 & -2 \\ 1 & 1 & -2 \\ 0 & 0 & 3 \end{bmatrix}, D = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -4 \end{bmatrix}$.

8.2.12: $P = \begin{bmatrix} 2 & 1 & 4 \\ -3 & 0 & -2 \\ 1 & 0 & 1 \end{bmatrix}, J = \begin{bmatrix} -4 & 1 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & -2 \end{bmatrix}$.

9.1.2: $\omega = \pi \sqrt{\frac{15}{2}}$

9.1.5: $\lambda_k = 4k^2\pi^2$ for $k = 1, 2, 3, \dots$ $x_k = \cos(2k\pi t) + B \sin(2k\pi t)$ (for any B)

9.1.8: $x(t) = -\sin(t)$

9.1.10: General solution is $x = Ce^{-\lambda t}$. Since $x(0) = 0$ then $C = 0$, and so $x(t) = 0$. Therefore, the solution is always identically zero. One condition is always enough to guarantee a unique solution for a first order equation.

9.1.11: $\frac{\sqrt{3}}{3}e^{\frac{-3}{2}\sqrt[3]{\lambda}} - \frac{\sqrt{3}}{3}\cos\left(\frac{\sqrt{3}\sqrt[3]{\lambda}}{2}\right) + \sin\left(\frac{\sqrt{3}\sqrt[3]{\lambda}}{2}\right) = 0$

9.2.3: $\sin(t)$

9.2.7: $\sum_{n=1}^{\infty} \frac{(\pi-n)\sin(\pi n+\pi^2)+(\pi+n)\sin(\pi n-\pi^2)}{\pi n^2-\pi^3} \sin(nt)$

9.2.10: $\frac{1}{2} - \frac{1}{2} \cos(2t)$

9.2.12: $\frac{\pi^4}{5} + \sum_{n=1}^{\infty} \frac{(-1)^n(8\pi^2 n^2 - 48)}{n^4} \cos(nt)$

9.3.4: a) $\frac{8}{6} + \sum_{n=1}^{\infty} \frac{16(-1)^n}{\pi^2 n^2} \cos\left(\frac{n\pi}{2}t\right)$ b) $\frac{8}{6} - \frac{16}{\pi^2} \cos\left(\frac{\pi}{2}t\right) + \frac{4}{\pi^2} \cos(\pi t) - \frac{16}{9\pi^2} \cos\left(\frac{3\pi}{2}t\right) + \dots$

9.3.6: a) $\sum_{n=1}^{\infty} \frac{(-1)^{n+1} 2\lambda}{n\pi} \sin\left(\frac{n\pi}{\lambda}t\right)$ b) $\frac{2\lambda}{\pi} \sin\left(\frac{\pi}{\lambda}t\right) - \frac{\lambda}{\pi} \sin\left(\frac{2\pi}{\lambda}t\right) + \frac{2\lambda}{3\pi} \sin\left(\frac{3\pi}{\lambda}t\right) - \dots$

9.3.8: $f'(t) = \sum_{n=1}^{\infty} \frac{\pi}{n^2+1} \cos(n\pi t)$

9.3.11: a) $F(t) = \frac{t}{2} + C + \sum_{n=1}^{\infty} \frac{1}{n^4} \sin(nt)$ b) no

9.3.15: a) $\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin(nt)$ b) f is continuous at $t = \pi/2$ so the Fourier series converges to $f(\pi/2) = \pi/4$. Obtain $\pi/4 = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{2n-1} = 1 - 1/3 + 1/5 - 1/7 + \dots$. c) Using the first 4 terms get $76/105 \approx 0.72$ (quite a bad approximation, you would have to take about 50 terms to start to get to within 0.01 of $\pi/4$).

9.3.17: a) $F(0) = 1$, b) $F(-1) = 0$, c) $F(1) = 2$, d) $F(-2) = 1$, e) $F(4) = 1$, f) $F(-9) = 0$

9.4.7: a) $\frac{1}{2} + \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{-4}{\pi^2 n^2} \cos\left(\frac{n\pi}{3}t\right)$ b) $\sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{\pi n} \sin\left(\frac{n\pi}{3}t\right)$

9.4.9: a) $\cos(2t)$ b) $\sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{-4n}{\pi n^2 - 4\pi} \sin(nt)$

9.4.10: a) $f(t)$ b) 0

9.4.13: $\sum_{n=1}^{\infty} \frac{-1}{n^2(1+n^2)} \sin(nt)$

9.4.16: $\frac{t}{\pi} + \sum_{n=1}^{\infty} \frac{1}{2^n(\pi-n^2)} \sin(nt)$

9.5.3: $x = \frac{1}{\sqrt{2}-4\pi^2} \sin(2\pi t) + \frac{0.1}{\sqrt{2}-100\pi^2} \cos(10\pi t)$

9.5.5: $x = \sum_{n=1}^{\infty} \frac{e^{-n}}{3-(2n)^2} \cos(2nt)$

9.5.8: $x = \frac{1}{2\sqrt{3}} + \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{-4}{n^2\pi^2(\sqrt{3}-n^2\pi^2)} \cos(n\pi t)$

9.5.10: $x = \frac{1}{2\sqrt{3}} - \frac{2}{\pi^3} t \sin(\pi t) + \sum_{\substack{n=3 \\ n \text{ odd}}}^{\infty} \frac{-4}{n^2\pi^4(1-n^2)} \cos(n\pi t)$

10.1.2: a) $u = \frac{1}{1+(x+5t)^2}$ b) $u = \cos(x-2t)$

10.1.4: $u = \cos(x-t)e^{-t^2/2}$

10.1.7: $u = x + 4t$

10.2.7: $u(x, t) = e^{\lambda t} e^{\lambda x}$ for some λ

10.2.8: $u(x, t) = Ae^x + Be^t$

10.3.3: $u(x, t) = 5 \sin(x) e^{-3t} + 2 \sin(5x) e^{-75t}$

10.3.5: $u(x, t) = 1 + 2 \cos(x) e^{-0.1t}$

10.3.12: a) 0, b) minimum -100 , maximum 100 , c) $t = \frac{\ln 2}{4\pi^2}$.

10.4.4: $y(x, t) = \sin(x)(\sin(t) + \cos(t))$

10.4.6: $y(x, t) = \frac{1}{5\pi} \sin(\pi x) \sin(5\pi t) + \frac{1}{100\pi} \sin(2\pi x) \sin(10\pi t)$

10.4.7: $y(x, t) = \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin(nx) \cos(n\sqrt{2}t)$

10.4.10: $y(x, t) = \sin(2x) + t \sin(x)$

10.5.3: $y(x, t) = \frac{\sin(2\pi(x-3t))+\sin(2\pi(3t+x))}{2} + \frac{\cos(3\pi(x-3t))-\cos(3\pi(3t+x))}{18\pi}$

10.5.6: a) $y(x, 0.1) = \begin{cases} x - x^2 - 0.04 & \text{if } 0.2 \leq x \leq 0.8 \\ 0.6x & \text{if } x \leq 0.2 \\ 0.6 - 0.6x & \text{if } x \geq 0.8 \end{cases}$

b) $y(x, 1/2) = -x + x^2$ c) $y(x, 1) = x - x^2$

10.5.9: a) $y(1, 1) = -1/2$ b) $y(4, 3) = 0$ c) $y(3, 9) = 1/2$

10.6.7: $u(x, y) = \sum_{n=1}^{\infty} \frac{1}{n^2} \sin(n\pi x) \left(\frac{\sinh(n\pi(1-y))}{\sinh(n\pi)} \right)$

10.6.11: $u(x, y) = 0.1 \sin(\pi x) \left(\frac{\sinh(\pi(2-y))}{\sinh(2\pi)} \right)$

10.7.3: $u = 1 + \sum_{n=1}^{\infty} \frac{1}{n^2} r^n \sin(n\theta)$

10.7.4: $u = 1 - x$

10.7.9: a) $u = \frac{-1}{4}r^2 + \frac{1}{4}$ b) $u = \frac{-1}{4}r^2 + \frac{1}{4} + r^2 \sin(2\theta)$

10.7.13: $u(r, \theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\rho^2 - r^2}{\rho - 2r\rho \cos(\theta - \alpha) + r^2} g(\alpha) d\alpha$